Human Computer Interaction: Concepts, Methodologies, Tools, and Applications

Panayiotis Zaphiris  
*City University of London, UK*

Chee Siang Ang  
*City University of London, UK*
Editor-in-Chief

Mehdi Khosrow-Pour, DBA
Editor-in-Chief
Contemporary Research in Information Science and Technology, Book Series

Associate Editors

Steve Clarke
University of Hull, UK

Murray E. Jennex
San Diego State University, USA

Annie Becker
Florida Institute of Technology USA

Ari-Veikko Anttiloiko
University of Tampere, Finland

Editorial Advisory Board

Sherif Kamel
American University in Cairo, Egypt

In Lee
Western Illinois University, USA

Jerzy Kisielnicki
Warsaw University, Poland

Keng Siau
University of Nebraska-Lincoln, USA

Amar Gupta
Arizona University, USA

Craig van Slyke
University of Central Florida, USA

John Wang
Montclair State University, USA

Vishanth Weerakkody
Brunel University, UK
Additional Research Collections found in the
“Contemporary Research in Information Science and Technology”
Book Series

Data Mining and Warehousing: Concepts, Methodologies, Tools, and Applications
John Wang, Montclair University, USA • 6-volume set • ISBN 978-1-60566-056-1

Electronic Business: Concepts, Methodologies, Tools, and Applications
In Lee, Western Illinois University • 4-volume set • ISBN 978-1-59904-943-4

Electronic Commerce: Concepts, Methodologies, Tools, and Applications
S. Ann Becker, Florida Institute of Technology, USA • 4-volume set • ISBN 978-1-59904-943-4

Electronic Government: Concepts, Methodologies, Tools, and Applications
Ari-Veikko Anttiroiko, University of Tampere, Finland • 6-volume set • ISBN 978-1-59904-947-2

Knowledge Management: Concepts, Methodologies, Tools, and Applications
Murray E. Jennex, San Diego State University, USA • 6-volume set • ISBN 978-1-59904-933-5

Information Communication Technologies: Concepts, Methodologies, Tools, and Applications
Craig Van Slyke, University of Central Florida, USA • 6-volume set • ISBN 978-1-59904-949-6

Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications
Vijayan Sugumaran, Oakland University, USA • 4-volume set • ISBN 978-1-59904-941-0

Information Security and Ethics: Concepts, Methodologies, Tools, and Applications

Medical Informatics: Concepts, Methodologies, Tools, and Applications
Joseph Tan, Wayne State University, USA • 4-volume set • ISBN 978-1-60566-050-9

Mobile Computing: Concepts, Methodologies, Tools, and Applications
David Taniar, Monash University, Australia • 6-volume set • ISBN 978-1-60566-054-7

Multimedia Technologies: Concepts, Methodologies, Tools, and Applications
Syed Mahbubur Rahman, Minnesota State University, Mankato, USA • 3-volume set • ISBN 978-1-60566-054-7

Virtual Technologies: Concepts, Methodologies, Tools, and Applications
Jerzy Kisielnicki, Warsaw University, Poland • 3-volume set • ISBN 978-1-59904-955-7

Free institution-wide online access with the purchase of a print collection!
List of Contributors

Aaltio, Iiris / Lappeenranta University of Technology, Finland .......................................................... 2030
Abraham, J. R. / University of the West of England, UK ................................................................. 376
Abutarbush, Sameeh M. / University of Prince Edward Island, Canada ........................................ 1656
Adam, Alison / University of Salford, UK ......................................................................................... 2435
Ágerfalk, Pär J. / University of Limerick, Ireland & Örebro University, Sweden ............................ 2341
Ahuja, Manju / Indiana University, USA ......................................................................................... 1736
Ala-Rämi, Katarina / University of Oulu, Finland ........................................................................... 2037
Albert, Sylvie / Laurentian University, Canada .................................................................................. 1915
Aldiri, Khalid / University of Bradford, UK ...................................................................................... 2235
Almeida, Hyggo / Federal University of Campina Grande, Brazil ................................................. 228, 254, 279
Almendros-Jimenez, Jesus M. / Universidad de Almeria, Spain ....................................................... 404
Alpert, Sherman R. / IBM T. J. Watson Research Center, USA ...................................................... 177
Altenhofen, Michael / SAP Research, CEC Karlsruhe, Germany .................................................... 2144
Anagnostopoulos, Christos / University of Athens, Greece ............................................................... 1096
Andersen, Peter Bøgh / University of Aarhus, Denmark ................................................................. 552
Andrade, Stephen C. / Johnson & Wales University, USA ............................................................... 2590
Andrea, Carugati / Catholic University of Lille, France ................................................................. 489
Anke, Juergen / Dresden University of Technology, Germany ......................................................... 168
Ardissono, Liliana / Università de Torino, Italy ................................................................................. 791
Avery Gomez, Elizabeth / New Jersey Institute of Technology, USA ........................................... 1334
Baber, Chris / The University of Birmingham, UK ........................................................................... 80, 1134
Barbosa, Nadia / Federal University of Campina Grande, Brazil .................................................... 279
Beckwith, Laura / Oregon State University, USA ............................................................................. 25
Bellotti, Francesco / University of Genoa, Italy ............................................................................... 1234
Ben Zakour, Amel / FSJEGJ, Tunisia ................................................................................................... 132
Bernsteiner, Reinhard / University for Health Sciences, Medical Informatics and Technology, Austria ................................................................. 1402
Berta, Riccardo / University of Genoa, Italy ..................................................................................... 1234
Bharati, Pratyush / University of Massachusetts, Boston, USA ...................................................... 1956
Bhattacherjee, Anol / University of South Florida, USA ................................................................. 1608
Bias, Randolph G. / The University of Texas at Austin, USA ......................................................... 2404
Bieber, Michael / New Jersey Institute of Technology, USA ............................................................ 1334
Blocher, J. Michael / Northern Arizona University, USA ............................................................... 1547
Blomqvist, Kirsi-Marja / Lappeenranta University of Technology, Finland ..................................... 2181
Bondarouk, Tanya / University of Twente, The Netherlands ............................................................ 2214
Bonk, Curtis J. / Indiana University, USA ......................................................... 1262
Bornhövd, Christof / SAP Research, Research Center Palo Alto, USA .................. 2144
Bourgue, Marie-Luce / Queen Mary University of London, UK .......................... 95
Bowen, Jonathan P. / London South Bank University, UK .............................. 1629
Boyle, Randall J. / University of Utah, USA ..................................................... 1970
Brey, Philip / University of Twente, The Netherlands ..................................... 154
Bruegger, Pascal / University of Fribourg, Switzerland .................................. 1015
Brynskov, Martin / Center for Interactive Spaces, University of Aarhus, Denmark .. 552
Bublitz, Frederico / Federal University of Campina Grande, Brazil ................. 279
Burnett, Gary / University of Nottingham, UK ............................................... 532
Burnett, Margaret / Oregon State University, USA ........................................ 25
Bushey, Robert G. / AT&T Laboratories, USA .............................................. 2404
Bygstad, Bendik / Norwegian School of Information Technology, Norway .......... 2090
Caronia, Letizia / University of Bologna, Italy ............................................. 2655
Castelfranchi, Cristiano / Institute for Cognitive Sciences and Technology, National Research Council of Italy, Italy ...................................................... 2316
Catarci, Tiziana / University of Rome “La Sapienza”, Italy ............................... 872, 928
Chaudhury, Adhijit / Bryant University, USA .............................................. 1956
Chen, Chaomei / Drexel University, USA ..................................................... 1678
Chen, Yen-Jung / National University of Tainan, Taiwan ............................... 1330
Cheuk, Bonnie Wai-yi / Improvement Service for the Scottish Local Government, Scotland ................................................................. 1905
Cigognini, M. E. / University of Florence (IT), Italy ...................................... 961
Clark, Chris / University of Saskatchewan, Canada ........................................ 1656
Cohn, Jonathan / UCLA, USA ................................................................. 2530
Compagna, Luca / SAP Research, France .................................................. 2424
Cox, Anna L. / University College London, UK ........................................... 1154
Crombie, David / DEDICON S.r.l., The Netherlands .................................... 1350
Cyr, Dianne / Simon Fraser University, Canada ........................................... 1761
da Cunha Borelli, Walter / State University of Campinas, Brazil ................. 1037
Dai, Honghua / DePaul University, USA .................................................... 732
Daniel, Ben K. / University of Saskatchewan, Canada ................................... 979
Das, Sajal / University of Texas at Arlington, USA ....................................... 1173
Davenport, Glorianna / The Media Laboratory, USA .................................. 1521
Davis, Donald D. / Old Dominion University, USA ...................................... 2023
de Fátima Queiroz Vieira Turnell, Maria / Universidade Federal de Campina Grande (UFCG), Brazil ................................................................. 2573
De Gloria, Alessandro / University of Genoa, Italy ...................................... 1234
de Queiroz, José Eustáquio Rangel / Universidade Federal de Campina Grande (UFCG), Brazil ................................................................. 2573
de Sousa Ferreira, Danilo / Universidade Federal de Campina Grande (UFCG), Brazil 2573
Deng, Li / Microsoft Research, USA .......................................................... 997
Deng, Liqiong / University of West Georgia, USA ........................................ 850
Diaper, Dan / Middlesex University, UK ..................................................... 70
Dix, Alan / Lancaster University, UK .......................................................... 928
Doyle, D. John / Cleveland Clinic Foundation, USA .................................... 483
Doyle, Linda / University of Dublin, Ireland ................................................................. 2266
Edenius, Mats / Stockholm School of Economics, Sweden ........................................... 2376
El Louadi, Mohamed / University of Tunis, Tunisia ................................................. 123
Elia, Hadzilias / Catholic University of Lille, France ............................................. 489
Erlendson, Benjamin E. / Arizona State University, USA ...................................... 698
Esteves, Jose / Universidade Politecnica de Catalunya, Barcelona, Spain ........ 2132
Evangelou, Christina E. / Research Academic Computer Technology Institute, Greece ... 1044
Falcone, Rino / Institute for Cognitive Sciences and Technology, National Research Council of Italy, Italy ................................................................. 2316
Fan, Haiyan / Texas A&M University, USA ............................................................ 850
Farmer, Rod / The University of Melbourne, Australia ........................................ 899
Ferreira, Glauber / Federal University of Campina Grande, Brazil .................... 228, 279
Filippini-Fantoni, Silvia / The University of Paris I Sorbonne University, France ...... 1629
Fischbach, Kai / University of Cologne, Germany ............................................... 2191
Fitzgerald, Robert / University of Canberra, Australia ........................................ 1459
Fletcher, Gordon / University of Salford, UK ...................................................... 1871
Floridi, Luciano / Università degli Studi di Bari, Italy and Oxford University, UK ...... 2450
Foth, Marcus / Queensland University of Technology, Australia .................. 589
French, Sheila / Manchester Metropolitan University, UK ................................ 2296
Fu, Guohong / The University of Hong Kong, Hong Kong ..................................... 1595
Furtado, Elizabeth / University of Fortaleza, Brazil ........................................... 448
Fuschi, David Luigi / GIUNTI Labs S.r.l., Italy .................................................. 1350
Fussell, Susan R. / Carnegie Mellon University, USA ......................................... 1801
Gabrielli, Silvia / University of Rome “La Sapienza”, Italy ...................................... 928
Gallagher, Kevin / Northern Kentucky University, USA ..................................... 2052
Gefen, David / Drexel University, USA ............................................................. 2644
Gena, Cristina / Università di Torino, Italy .......................................................... 791
George, Joey F. / Florida State University, USA .................................................. 1970
Georgiadis, Christos K. / University of Macedonia, Thessaloniki, Greece ............ 393
Geri, Nitza / The Open University of Israel, Israel ............................................. 2644
Gkotsis, George / Research Academic Computer Technology Institute, Greece .... 1044
Gold, Stuart S. / DeVry University, USA ............................................................. 1864
Goldkuhl, Göran / Linköping University, Sweden & Jönköping International Business School, Sweden .......................................................... 2071, 2341
Greenhill, Anita / The University of Manchester, UK .......................................... 1871, 2259
Greitzer, Frank L. / Pacific Northwest National Laboratory, USA .................. 2693
Griffith, Douglas / General Dynamics Advanced Information Systems, USA ..... 2693
Gruba, Paul / The University of Melbourne, Australia ........................................ 899
Guo, Huaqun / Institute for Infocomm Research and National University of Singapore, Singapore .......................................................... 1087
Gurevych, Iryna / Technische Universität Darmstadt, Germany ......................... 1
Gurram, Rama / SAP Research, USA ................................................................. 102
Hadjithymiades, Stathes / University of Athens, Greece ..................................... 1096
Hall, Jon G. / The Open University, UK ............................................................. 713
Hanrahan, Michael / Bates College, USA ............................................................ 1251
Heinemann, Andreas / Technical University of Darmstadt, Germany .............. 333
Henttonen, Kaisa / Lappeenranta University of Technology, Finland .............................................. 2181
Her, One-Soon / National Chengchi University, Taiwan ................................................................. 2499
Herdin, Thomas / University of Salzburg, Austria ........................................................................ 1055
Herring, Susan C. / Indiana University, USA .................................................................................. 1736
Hirsbrunner, Béat / University of Fribourg, Switzerland ............................................................... 1015
Hobbs, Dave / University of Bradford, UK .................................................................................... 2235
Hofkirchner, Wolfgang / University of Salzburg, Austria .............................................................. 1055
Hohmann, Claus / Bergische Universität Wuppertal, Germany ....................................................... 2462
Hornik, Steven / University of Central Florida, USA .................................................................. 1885
Horowitz, Ellis / University of Southern California, USA .............................................................. 659
Hsu, HY Sonya / Southern Illinois University, USA ....................................................................... 1649
Hsu, Jeffrey / Fairleigh Dickinson University, USA ......................................................................... 1308
Hsu, Shang Hwa / National Chiao Tung University, Taiwan ......................................................... 516
Huang, Haiyan / The Pennsylvania State University, USA ............................................................ 1753
Huang, Rui / University of Texas at Arlington, USA .................................................................... 1173
Hubona, Geoffrey S. / Georgia State University, USA ................................................................. 2473
Huffaker, David A. / Northwestern University, USA ..................................................................... 1484
Hürt, Wolfgang / Albert-Ludwigs-Universität Freiburg, Germany .............................................. 1436
Hwang, Gwo-Jen / National University of Tainan, Taiwan ............................................................. 1330
Iribarne, Luis / Universidad de Almeria, Spain ............................................................................. 404
Isomäki, Hannakaisa / University of Lapland, Finland ................................................................. 605
James, Frankie / SAP Research, USA ............................................................................................ 102
Jaspers, M.W.M. / University of Amsterdam, The Netherlands .................................................. 622
Jefsioutine, Marie / University of Central England - Birmingham, UK ....................................... 432
Jin, Dal Yong / Simon Fraser University, Canada ........................................................................... 1200
José Luzón, Maria / University of Zaragoza, Spain ....................................................................... 1291
Julien, Christine / The University of Texas at Austin, USA ............................................................. 2708
Kabadayi, Sanem / The University of Texas at Austin, USA .......................................................... 2708
Kacmar, Charles J. / The University of Alabama, USA ................................................................. 1970
Karali, Izambo / University of Athens, Greece ............................................................................. 1096
Karousos, Nikos / National Hellenic Research Foundation, Greece ........................................... 1044
Khattab, Ishragna / Brunel University, UK ..................................................................................... 1569
Kidd, Terry T. / University of Texas School of Public Health, USA ............................................. 1468
Kimani, Stephen / Jomo Kenyatta University of Agriculture and Technology, Kenya, & University of Rome “La Sapienza”, Italy ......................................................... 872, 928
Kirakowski, Jurek / University College Cork, Ireland ................................................................. 361
Klassen, Myungsook / California Lutheran University, USA ......................................................... 1729
Kleinschmidt, João Henrique / State University of Campinas, Brazil ......................................... 1037
Klinker, Gudrun / Technische Universität München, Germany ...................................................... 629
Klobas, Jane / Bocconi University, Italy & University of Western Australia, Australia ................ 1283
Knight, James / The University of Birmingham, UK ...................................................................... 1134
Knight, John / University of Central England - Birmingham, UK ............................................. 432
Knott, Benjamin A. / U.S. Air Force Research Laboratory, USA .................................................. 2404
Kolomvatsos, Kostas / National & Kapodistrian University of Athens, Greece ......................... 1503
Kortum, Kortum / Rice University, USA ....................................................................................... 2404
<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muhle, Steffen</td>
<td>University of Cologne, Germany</td>
</tr>
<tr>
<td>Mühlhäuser, Max</td>
<td>Technische Universität Darmstadt, Germany</td>
</tr>
<tr>
<td>Murphy, Elizabeth D.</td>
<td>U.S. Census Bureau, USA</td>
</tr>
<tr>
<td>Naylor, Jonathan M.</td>
<td>Ross University School of Veterinary Medicine, West Indies</td>
</tr>
<tr>
<td>Ngoh, Lek-Heng</td>
<td>Institute for Infocomm Research, A*STAR, Singapore</td>
</tr>
<tr>
<td>Nousia, Dora</td>
<td>Research Academic Computer Technology Institute, Greece</td>
</tr>
<tr>
<td>Numerico, Teresa</td>
<td>London South Bank University, UK</td>
</tr>
<tr>
<td>O’Brien, David</td>
<td>University of Saskatchewan, Canada</td>
</tr>
<tr>
<td>O’Connell, Theresa A.</td>
<td>Humans &amp; Computers, Inc., USA</td>
</tr>
<tr>
<td>O’Hare, Gregory M. P.</td>
<td>University College Dublin, Ireland</td>
</tr>
<tr>
<td>Oberle, Daniel</td>
<td>SAP Research, CEC Karlsruhe, Germany</td>
</tr>
<tr>
<td>Ocker, Rosalie J.</td>
<td>The Pennsylvania State University, USA</td>
</tr>
<tr>
<td>Ogan, Christine</td>
<td>Indiana University, USA</td>
</tr>
<tr>
<td>O’Grady, Michael J.</td>
<td>University College Dublin, Ireland</td>
</tr>
<tr>
<td>Olaniran, Bolanle A.</td>
<td>Texas Tech University, USA</td>
</tr>
<tr>
<td>Oliveira, Lorenzo</td>
<td>Federal University of Campina Grande, Brazil</td>
</tr>
<tr>
<td>Ong, Bee</td>
<td>University of Leeds, UK</td>
</tr>
<tr>
<td>Ostermann, Herwig</td>
<td>University for Health Sciences, Medical Informatics and Technology, Austria</td>
</tr>
<tr>
<td>Oulasvirta, Antti</td>
<td>Helsinki Institute for Information Technology, Finland</td>
</tr>
<tr>
<td>Owens, Emiel</td>
<td>Texas Southern University, USA</td>
</tr>
<tr>
<td>Pallotta, Vincenzo</td>
<td>University of Fribourg, Switzerland</td>
</tr>
<tr>
<td>Panayiotaki, Angeliki</td>
<td>University of Patras, Greece</td>
</tr>
<tr>
<td>Paravastu, Narasimha</td>
<td>Central State University, USA</td>
</tr>
<tr>
<td>Parchoma, Gale</td>
<td>University of Saskatchewan, Canada</td>
</tr>
<tr>
<td>Parmee, I. C.</td>
<td>University of the West of England, UK</td>
</tr>
<tr>
<td>Passerini, Katia</td>
<td>New Jersey Institute of Technology, USA</td>
</tr>
<tr>
<td>Pastor, Joan</td>
<td>Universitat Internacional de Catalunya, Barcelona, Spain</td>
</tr>
<tr>
<td>Pavlovski, Christopher J.</td>
<td>IBM Corporation, Australia</td>
</tr>
<tr>
<td>Pearson, Will</td>
<td>Sheffield Hallam University, UK</td>
</tr>
<tr>
<td>Peebles, David</td>
<td>University of Huddersfield, UK</td>
</tr>
<tr>
<td>Perkusich, Angelo</td>
<td>Federal University of Campina Grande, Brazil</td>
</tr>
<tr>
<td>Pettenati, M. C.</td>
<td>University of Florence (IT), Italy</td>
</tr>
<tr>
<td>Polley, Lyddden</td>
<td>University of Saskatchewan, Canada</td>
</tr>
<tr>
<td>Polovina, Simon</td>
<td>Sheffield Hallam University, UK</td>
</tr>
<tr>
<td>Porterfield, Sharon</td>
<td>University of Saskatchewan, Canada</td>
</tr>
<tr>
<td>Qahwaji, Rami</td>
<td>University of Bradford, UK</td>
</tr>
<tr>
<td>Quah, Jon T. S.</td>
<td>Nanyang Technological University, Singapore</td>
</tr>
<tr>
<td>Rahman, Hakikur</td>
<td>SDNP Bangladesh, Bangladesh</td>
</tr>
<tr>
<td>Rahman, Lutfar</td>
<td>Association for Advancement of Information Technology (AAIT), Bangladesh</td>
</tr>
<tr>
<td>Rahman, Nusrat</td>
<td>University of Dhak and Association</td>
</tr>
<tr>
<td>Rapanotti, Lucia</td>
<td>The Open University, UK</td>
</tr>
<tr>
<td>Reddick, Christopher G.</td>
<td>The University of Texas at San Antonio, USA</td>
</tr>
</tbody>
</table>
Richter, Kai / Computer Graphics Centre (ZGDV), Germany .......................................................... 323
Rigou, Maria / University of Patras, Greece & Research Acaademic Computer Technology
Institute, Greece ......................................................................................................................... 992, 2164
Rittgen, Peter / University College of Borås, Sweden .................................................................. 2390
Robinson, Jean C. / Indiana University, USA ............................................................................... 1736
Roth, Volker / OGM Laboratory LLC, USA .................................................................................. 323
Ruzic, Fjodor / Institute for Informatics, Croatia ......................................................................... 2324
Saddler, Paula F. / Association of Certified Fraud Examiners, USA ........................................... 2023
Sajjapanroj, Suthiporn / Indiana University, USA ...................................................................... 1262
Salah Hamdi, Mohamed / University of Qatar, Qatar .................................................................. 1118
Salovaara, Antti / Helsinki Institute for Information Technology, Finland .................................. 20
Sánchez-Franco, Manuel Jesús / University of Seville, Spain ...................................................... 1995
Sandberg, Karl W. / Mid Sweden University, Sweden ................................................................. 60
Sandor, Christian / Technische Universität München, Germany .................................................. 629
Sanford, Clive / Aalborg University, Denmark ............................................................................ 1608
Sarkar, Asit / University of Saskatchewan, Canada ..................................................................... 979
Sater Carstens, Deborah / Florida Institute of Technology, USA ............................................... 1769
Sawasdichai, Napawan / King Mongkut’s Institute of Technology, Thailand ............................... 816
Schilling, Albert / University of Fortaleza, Brazil ......................................................................... 448
Schmitt, Christian / University of Cologne, Germany ................................................................. 2191
Schoeder, Detlef / University of Cologne, Germany ................................................................. 2191
Schwarz, Kathy / University of Saskatchewan, Canada ............................................................... 1656
Selgrade, Katherine A. / Old Dominion University, USA ............................................................ 2023
Seok, Soonhwa / The University of Kansas, USA ....................................................................... 2542
Serenko, Alexander / Lakehead University, Canada .................................................................... 1069
Serour, Magdy K. / University of Technology, Sydney, Australia ............................................. 1937
Seta, Kazuhiisa / Osaka Prefecture University, Japan .................................................................... 950
Setlock, Leslie D. / Carnegie Mellon University, USA ................................................................. 1801
Shah, Abad / R & D Center of Computer Science, Pakistan ....................................................... 659
Sharma, Ramesh C. / Indira Gandhi National Open University, India ......................................... 1777
Sharples, Mike / University of Nottingham, Jubilee Campus, UK ............................................. 677
Shirah, Gregory W. / National Aeronautics and Space Administration, USA .............................. 2473
Shmon, Cindy L. / University of Saskatchewan, Canada ............................................................... 1656
Siau, Keng / University of Nebraska-Lincoln, USA .................................................................. 645
Singhera, Zafar / ZAF Consulting, USA ..................................................................................... 659
Sirmakessis, Spiros / Technological Institution of Messolongi, Greece & Research
Academic Computer Technology Institute, Greece .......................................................... 992, 2164
Song, Holim / Texas Southern University, USA .......................................................................... 1468
Sorte, Shraddha / Oregon State University, USA ........................................................................ 25
Sousa, Kenia / University of Fortaleza, Brazil ........................................................................... 448
Spedding, Paul / University of Salford, UK .............................................................................. 2435
Staudinger, Roland / University for Health Sciences, Medical Informatics
and Technology, Austria ........................................................................................................ 1402
Stockard, Russell / California Lutheran University, USA ............................................................ 1729
Straub, Tobias / Fraunhofer Institute for Secure Information Technology, Germany ............. 333
Sundaram, David / University of Auckland, New Zealand ........................................................ 168
Contents
by Volume

Volume I.

Section I. Fundamental Concepts and Theories

This section serves as the foundation for this exhaustive reference tool by addressing crucial theories essential to the understanding of human-computer interaction. Chapters found within these pages provide an excellent framework in which to position human-computer interaction within the field of information science and technology. Individual contributions provide overviews of ubiquitous computing, cognitive informatics, and sociotechnical theory, while also exploring critical stumbling blocks of this field. Within this introductory section, the reader can learn and choose from a compendium of expert research on the elemental theories underscoring the research and application of human-computer interaction.

Chapter 1.1 Introduction to Ubiquitous Computing /
Max Mühlhäuser, Technische Universität Darmstadt, Germany;
and Iryna Gurevych, Technische Universität Darmstadt, Germany ............................................................... 1

Chapter 1.2 Ubiquitous Computing and the Concept of Context /
Antti Oulasvirta, Helsinki Institute for Information Technology, Finland;
and Antti Salovaara, Helsinki Institute for Information Technology, Finland ................................................. 20

Chapter 1.3 Gender and End-User Computing / Laura Beckwith, Oregon State University, USA;
Margaret Burnett, Oregon State University, USA;
and Shraddha Sorte, Oregon State University, USA ......................................................................................... 25

Chapter 1.4 The Theoretical Framework of Cognitive Informatics /
Yingxu Wang, University of Calgary, Canada ..................................................................................................... 33

Chapter 1.5 Human Factors in Public Information Systems /
Karl W. Sandberg, Mid Sweden University, Sweden;
and Pan Yan, Luleå University of Technology, Sweden ...................................................................................... 60
Chapter 1.6 Task Analysis at the Heart of Human-Computer Interaction / 
Dan Diaper, Middlesex University, UK ................................................................. 70

Chapter 1.7 Evaluating Mobile Human-Computer Interaction / 
Chris Baber, The University of Birmingham, UK .................................................. 80

Chapter 1.8 An Overview of Multimodal Interaction Techniques and Applications / 
Marie-Luce Bourguet, Queen Mary University of London, UK ................................. 95

Chapter 1.9 Multimodal and Federated Interaction / Frankie James, SAP Research, USA; and Rama Gurram, SAP Research, USA ........................................................................... 102

Chapter 1.10 The Arab World, Culture and Information Technology / 
Mohamed El Louadi, University of Tunis, Tunisia .................................................... 123

Chapter 1.11 Information Technology Acceptance across Cultures / 
Amel Ben Zakour, FSJEGJ, Tunisia ........................................................................... 132

Chapter 1.12 Is Information Ethics Culture-Relative? / 
Philip Brey, University of Twente, The Netherlands ............................................... 154

Chapter 1.13 Personalization Techniques and Their Application / 
Juergen Anke, Dresden University of Technology, Germany; and David Sundaram, University of Auckland, New Zealand ................................................................. 168

Chapter 1.14 User-Centered Evaluation of Personalized Web Sites: What’s Unique? / 
Sherman R. Alpert, IBM T. J. Watson Research Center, USA; and John G. Vergo, IBM T. J. Watson Research Center, USA ........................................................................... 177

Chapter 1.15 The Usability Engineering Behind User-Centered Processes 
for Web Site Development Lifecycles / Theresa A. O’Connell, Humans & Computers, Inc., USA; and Elizabeth D. Murphy, U.S. Census Bureau, USA ......................................................... 192

Chapter 1.16 Personalization Systems and Their Deployment as Web Site Interface 
Design Decisions / Nanda Kumar, Baruch College, The City University of New York, USA .............. 212

Chapter 1.17 Communication + Dynamic Interface = Better User Experience / 
Simon Polovina, Sheffield Hallam University, UK; and Will Pearson, Sheffield Hallam 
University, UK ........................................................................................................ 220

Chapter 1.18 Pervasive Computing: What Is It Anyway? / 
Emerson Loureiro, Federal University of Campina Grande, Brazil; 
Glauber Ferreira, Federal University of Campina Grande, Brazil; 
Hyggo Almeida, Federal University of Campina Grande, Brazil; and Angelo Perkusich, Federal University of Campina Grande, Brazil ......................................................... 228
Chapter 1.19 Pervasive and Grid Computing Merging / Loreno Oliveira, Federal University of Campina Grande, Brazil; Emerson Loureiro, Federal University of Campina Grande, Brazil; Higgo Almeida, Federal University of Campina Grande, Brazil; and Angelo Perkusich, Federal University of Campina Grande, Brazil ................................................................. 254

Chapter 1.20 Leveraging Pervasive and Ubiquitous Service Computing / Zhijun Zhang, University of Phoenix, USA ......................................................................................... 262

Chapter 1.21 Service Provision for Pervasive Computing Environments / Emerson Loureiro, Federal University of Campina Grande, Brazil; Frederico Bublitz, Federal University of Campina Grande, Brazil; Loreno Oliveira, Federal University of Campina Grande, Brazil; Nadia Barbosa, Federal University of Campina Grande, Brazil; Angelo Perkusich, Federal University of Campina Grande, Brazil; Higgo Almeida, Federal University of Campina Grande, Brazil; and Glauber Ferreira, Federal University of Campina Grande, Brazil ................................................................. 279


Chapter 1.23 Computer-Mediated Communication Research / J. D. Wallace, Lubbock Christian University, USA ................................................................................................. 299

Chapter 1.24 Computer-Mediated Communication in Virtual Learning Communities / Lisa Link, Flensburg University of Applied Sciences, Germany; and Daniela Wagner, University of Hildesheim, Germany ................................................................................... 316

Chapter 1.25 Human-Computer Interaction and Security / Kai Richter, Computer Graphics Centre (ZGDV), Germany; and Volker Roth, OGM Laboratory LLC, USA ........................................................................ 323

Chapter 1.26 Security for Ubiquitous Computing / Tobias Straub, Fraunhofer Institute for Secure Information Technology, Germany; and Andreas Heinemann, Technical University of Darmstadt, Germany ........................................................................ 333

Section II. Development and Design Methodologies

This section provides in-depth coverage of conceptual architectures, frameworks and methodologies related to the design and implementation of interfaces and technologies that promote successful human-computer interaction. Throughout these contributions, research fundamentals in the discipline are presented and discussed. From broad examinations to specific discussions on particular frameworks and infrastructures, the research found within this section spans the discipline while also offering detailed, specific discussions. Basic designs, as well as abstract developments, are explained within these chapters, and frameworks for designing and personalizing interfaces for medical equipment, televisions, and automobiles are discussed.
Chapter 2.1 Measuring the Human Element in Complex Technologies / Niamh McNamara, University College Cork, Ireland; and Jurek Kirakowski, University College Cork, Ireland ......................................................... 361

Chapter 2.2 Human-Centric Evolutionary Systems in Design and Decision-Making / I. C. Parmee, University of the West of England, UK; J. R. Abraham, University of the West of England, UK; and A. Machwe, University of the West of England, UK ............................................................................. 376

Chapter 2.3 Adaptation and Personalization of User Interface and Content / Christos K. Georgiadis, University of Macedonia, Thessaloniki, Greece ................................................................. 393

Chapter 2.4 User Interaction and Interface Design with UML / Jesus M. Almendros-Jimenez, Universidad de Almeria, Spain; and Luis Iribarne, Universidad de Almeria, Spain ........................................................................... 404

Chapter 2.5 Design Methods for Experience Design / Marie Jefsioutine, University of Central England–Birmingham, UK; and John Knight, University of Central England–Birmingham, UK ............................................................ 432

Chapter 2.6 Integrating Usability, Semiotic, and Software Engineering into a Method for Evaluating User Interfaces / Kenia Sousa, University of Fortaleza, Brazil; Albert Schilling, University of Fortaleza, Brazil; and Elizabeth Furtado, University of Fortaleza, Brazil ........................................................................... 448

Chapter 2.7 Guidance in the Interface: Effects of Externalizing Information During Problem Solving / Christof van Nimwegen, Utrecht University, The Netherlands; Hermina Tabachneck-Schijf, Utrecht University, The Netherlands; and Herre van Oostendorp, Utrecht University, The Netherlands ....................................................................... 465

Chapter 2.8 Ergonomic User Interface Design in Computerized Medical Equipment / D. John Doyle, Cleveland Clinic Foundation, USA ............................................................................................. 483

Chapter 2.9 Development Of E-Government Services For Cultural Heritage: Examining the Key Dimensions / Carugati Andrea, Catholic University of Lille, France; and Hadziliyas Elias, Catholic University of Lille, France .............................................................................................. 489

Chapter 2.10 An Activity-Oriented Approach to Designing a User Interface for Digital Television / Shang Hwa Hsu, National Chiao Tung University, Taiwan; Ming-Hui Weng, National Chiao Tung University, Taiwan; and Cha-Hoang Lee, National Chiao Tung University, Taiwan .............................................................................. 516

Chapter 2.11 Designing and Evaluating In-Car User-Interfaces / Gary Burnett, University of Nottingham, UK ......................................................................................................................... 532
Chapter 2.12 The Semiotics of Smart Appliances and Pervasive Computing / 
Peter Bøgh Andersen, University of Aarhus, Denmark; 
and Martin Brynskov, Center for Interactive Spaces, University of Aarhus, Denmark .................. 552

Chapter 2.13 Analyzing the Factors Influencing the Successful Design 
and Uptake of Interactive Systems to Support Social Networks in Urban Neighborhoods / 
Marcus Foth, Queensland University of Technology, Australia ......................................................... 589

Chapter 2.14 Different Levels of Information Systems Designers’ 
Forms of Thought and Potential for Human-Centered Design / 
Hannakaisa Isomäki, University of Lapland, Finland ........................................................................ 605

Chapter 2.15 The Think Aloud Method and User Interface Design / 
M.W.M. Jaspers, University of Amsterdam, The Netherlands ............................................................ 622

Chapter 2.16 Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces / 
Christian Sandor, Technische Universität München, Germany; 
and Gudrun Klinker, Technische Universität München, Germany ...................................................... 629

Chapter 2.17 Social Network Structures in Open Source Software Development Teams / 
Yuan Long, Colorado State University-Pueblo, USA; 
and Keng Siau, University of Nebraska-Lincoln, USA ....................................................................... 645

Volume II.

Chapter 2.18 A Graphical User Interface (GUI) Testing Methodology / 
Zafar Singhera, ZAF Consulting, USA; Ellis Horowitz, University of Southern California, USA; 
and Abad Shah, R & D Center of Computer Science, Pakistan ............................................................ 659

Chapter 2.19 Socio-Cognitive Engineering / 
Mike Sharples, University of Nottingham, Jubilee Campus, UK .......................................................... 677

Chapter 2.20 On the Cognitive Processes of Human Perception 
with Emotions, Motivations, and Attitudes / Yingxu Wang, University of Calgary, Canada .............. 685

Chapter 2.21 Sociotechnical System Design for Learning: Bridging 
the Digital Divide with CompILE / Benjamin E. Erlandson, Arizona State University, USA ............ 698

Chapter 2.22 Problem Frames for Sociotechnical Systems / Jon G. Hall, The Open University, UK; 
and Lucia Rapanotti, The Open University, UK ...................................................................................... 713

Chapter 2.23 Integrating Semantic Knowledge with Web Usage Mining for Personalization / 
Honghua Dai, DePaul University, USA; and Bamshad Mobasher, DePaul University, USA ............. 732
Chapter 2.24 Modeling Variant User Interfaces for Web-Based Software Product Lines / Suet Chun Lee, BUSINEX, Inc., USA ................................................................. 760

Chapter 2.25 A User-Centered Approach to the Retrieval of Information in an Adaptive Web Site / Cristina Gena, Università di Torino, Italy; and Liliana Ardisono, Università di Torino, Italy .... 791

Chapter 2.26 Auto-Personalization Web Pages / Jon T. S. Quah, Nanyang Technological University, Singapore; Winnie C. H. Leow, Singapore Polytechnic, Singapore; K. L. Yong, Nanyang Technological University, Singapore .................................................. 807

Chapter 2.27 A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites: A User-Centered Approach to Web Site Development / Napawan Sawasdichai, King Mongkut’s Institute of Technology, Thailand ............................................. 816

Chapter 2.28 Developing and Validating a Measure of Web Personalization Strategy / Haiyan Fan, Texas A&M University, USA; and Liqiong Deng, University of West Georgia, USA... 850

Chapter 2.29 User Interface Formalization in Visual Data Mining / Tiziana Catarci, University of Rome “La Sapienza”, Italy; Stephen Kimani, University of Rome “La Sapienza”, Italy; and Stefano Lodi, University of Bologna, Italy .......................................................... 872

Chapter 2.30 Understanding the Nature of Task Analysis in Web Design / Rod Farmer, The University of Melbourne, Australia; and Paul Gruba, The University of Melbourne, Australia ......................................................... 899

Chapter 2.31 Designing for Tasks in Ubiquitous Computing: Challenges and Considerations / Stephen Kimani, Jomo Kenyatta University of Agriculture and Technology, Kenya; Silvia Gabrielli, University of Rome “La Sapienza”, Italy; Tiziana Catarci, University of Rome “La Sapienza”, Italy; and Alan Dix, Lancaster University, UK ........................................................................ 928

Chapter 2.32 Task Ontology-Based Human-Computer Interaction / Kazuhisa Seta, Osaka Prefecture University, Japan ................................................................. 950

Chapter 2.33 Social Networking Theories and Tools to Support Connectivist Learning Activities / M. C. Pettenati, University of Florence (IT), Italy; and M. E. Cigognini, University of Florence (IT), Italy ............................................................... 961

Chapter 2.34 User-Centered Design Principles for Online Learning Communities: A Sociotechnical Approach for the Design of a Distributed Community of Practice / Ben K. Daniel, University of Saskatchewan, Canada; David O’Brien, University of Saskatchewan, Canada; and Asit Sarkar, University of Saskatchewan, Canada ........................................................................ 979
Chapter 2.35 Content Personalization for Mobile Interfaces /  
Spiridoula Koukia, University of Patras, Greece;  
Maria Rigou, University of Patras, Greece & Research Academic Computer Technology Institute, Greece; and Spiros Sirmakessis, Technological Institution of Messolongi, Greece  
Research Academic Computer Technology Institute, Greece

Chapter 2.36 Speech-Centric Multimodal User Interface Design in Mobile Technology /  
Dong Yu, Microsoft Research, USA; and Li Deng, Microsoft Research, USA

Chapter 2.37 Kinetic User Interfaces: Physical Embodied Interaction with Mobile Ubiquitous Computing Systems /  
Vincenzo Pallotta, University of Fribourg, Switzerland;  
Pascal Bruegger, University of Fribourg, Switzerland; and Béat Hirsbrunner, University of Fribourg, Switzerland

Section III. Tools and Technologies

This section presents extensive coverage of specific tools and technologies that humans interact with and react to in their daily lives. These chapters provide an in-depth analysis of devices and tools such as portable music players, mobile phones, and even blogs. Within these rigorously researched chapters, readers are presented with countless examples of the technologies that support and encourage societal development and their resulting impact.

Chapter 3.1 Enabling Technologies for Pervasive Computing /  
João Henrique Kleinschmidt, State University of Campinas, Brazil; and Walter da Cunha Borelli, State University of Campinas, Brazil

Chapter 3.2 Augmenting Collaboration with Personalization Services /  
Christina E. Evangelou, Research Academic Computer Technology Institute, Greece;  
Manolis Tzagarakis, Research Academic Computer Technology Institute, Greece;  
Nikos Karousos, Research Academic Computer Technology Institute, Greece;  
George Gkotsis, Research Academic Computer Technology Institute, Greece; and Dora Nousia, Research Academic Computer Technology Institute, Greece

Chapter 3.3 Culture and Technology: A Mutual-Shaping Approach /  
Thomas Herdin, University of Salzburg, Austria;  
Wolfgang Hofkirchner, University of Salzburg, Austria; and Ursula Maier-Rabler, University of Salzburg, Austria

Chapter 3.4 Importance of Interface Agent Characteristics from End-User Perspective /  
Alexander Serenko, Lakehead University, Canada

Chapter 3.5 Web Mining for Public E-Services Personalization /  
Penelope Markellou, University of Patras, Greece; Angeliki Panayiotaki, University of Patras, Greece; and Athanasios Tsakalidis, University of Patras, Greece
Chapter 3.16 Plagiarism, Instruction, and Blogs / Michael Hanrahan, Bates College, USA .......... 1251

Chapter 3.17 Twin Wiki Wonders? Wikipedia and Wikibooks as Powerful Tools for Online Collaborative Writing / Meng-Fen Grace Lin, University of Houston, USA; Curtis J. Bonk, Indiana University, USA; and Suthiporn Sajjanaraj, Indiana University, USA .... 1262

Chapter 3.18 Wikis as Tools for Collaboration / Jane Klobas, Bocconi University, Italy & University of Western Australia, Australia ................................................................. 1283

Chapter 3.19 Academic Weblogs as Tools for E-Collaboration Among Researchers / María José Luzón, University of Zaragoza, Spain ................................................................. 1291

Chapter 3.20 Assessing Weblogs as Education Portals / Ian Weber, Texas A&M University, USA ......................................................................................................................... 1298

Chapter 3.21 Innovative Technologies for Education and Learning: Education and Knowledge-Oriented Applications of Blogs, Wikis, Podcasts, and More / Jeffrey Hsu, Fairleigh Dickinson University, USA ......................................................................................... 1308

Chapter 3.22 Ubiquitous Computing Technologies in Education / Gwo-Jen Hwang, National University of Tainan, Taiwan; Ting-Ting Wu, National University of Tainan, Taiwan; and Yen-Jung Chen, National University of Tainan, Taiwan ........................................ 1330

Chapter 3.23 Utilizing Web Tools for Computer-Mediated Communication to Enhance Team-Based Learning / Elizabeth Avery Gomez, New Jersey Institute of Technology, USA; Dezhi Wu, Southern Utah University, USA; Katia Passerini, New Jersey Institute of Technology, USA; and Michael Bieber, New Jersey Institute of Technology, USA ......................................................................................... 1334

Volume III.

Chapter 3.24 How Technology Can Support Culture and Learning / David Luigi Fuschi, GIUNTI Labs S.r.l., Italy; Bee Ong, University of Leeds, UK; and David Crombie, DEDICON S.r.l., The Netherlands ......................................................... 1350

Chapter 3.25 Tangible User Interfaces as Mediating Tools within Adaptive Educational Environments / Daria Loi, RMIT University, Australia ......................................................................................... 1388

Chapter 3.26 Facilitating E-Learning with Social Software: Attitudes and Usage from the Student’s Point of View / Reinhard Bernsteiner, University for Health Sciences, Medical Informatics and Technology, Austria; Herwig Ostermann, University for Health Sciences, Medical Informatics and Technology, Austria; and Roland Staudinger, University for Health Sciences, Medical Informatics and Technology, Austria ......................................................................................... 1402
Section IV. Utilization and Application

This section introduces and discusses the ways in which particular technologies have impacted humanity and proposes new ways in which IT-related innovations can be implemented within organizations and in society as a whole. These particular selections highlight, among other topics, ubiquitous computing applications in an educational setting and computer-mediated communication. Contributions included in this section provide excellent coverage of today’s world and insight into how our interaction with technology impacts the fabric of our present-day global village.

Chapter 4.1 Human Computer Interaction and the Best Mix of Face-to-Face and E-Interactions in Educational Settings / Bolanle A. Olaniran, Texas Tech University, USA ................................. 1423

Chapter 4.2 Human Computer Interaction for Computer-Based Classroom Teaching / Wolfgang Hürst, Albert-Ludwigs-Universität Freiburg, Germany; and Khaireel A. Mohamed, Albert-Ludwigs-Universität Freiburg, Germany ........................................ 1436

Chapter 4.3 Wikis as an Exemplary Model of Open Source Learning / Robert Fitzgerald, University of Canberra, Australia ........................................................................................................ 1459

Chapter 4.4 Re-Examining the Socioeconomic Factors Affecting Technology Use in Mathematics Classroom Practices / Emiel Owens, Texas Southern University, USA; Holim Song, Texas Southern University, USA; and Terry T. Kidd, University of Texas School of Public Health, USA .................................................................................. 1468

Chapter 4.5 Let Them Blog: Using Weblogs to Advance Literacy in the K-12 Classroom / David A. Huffaker, Northwestern University, USA .................................................................................. 1484

Chapter 4.6 Ubiquitous Computing Applications in Education / Kostas Kolomvatsos, National & Kapodistrian University of Athens, Greece .............................................................. 1503

Chapter 4.7 Unraveling the Taste Fabric of Social Networks / Hugo Liu, The Media Laboratory, USA; Pattie Maes, The Media Laboratory, USA; and Glorianna Davenport, The Media Laboratory, USA .................................................................................. 1521

Chapter 4.8 Computer-Mediated Communication: Enhancing Online Group Interactions / J. Michael Blocher, Northern Arizona University, USA .............................................................................. 1547

Chapter 4.9 Mobile Phone Use Across Cultures: A Comparison Between the United Kingdom and Sudan / Ishraga Khattab, Brunel University, UK; and Steve Love, Brunel University, UK .................................................................................. 1569

Chapter 4.10 Gender and Computing at University in the UK / Ruth Woodfield, University of Sussex, UK .......................................................................................................................... 1583

Chapter 4.11 Gender and Information Technology in Rural Bangladesh / Lutfor Rahman, Association for Advancement of Information Technology (AAIT), Bangladesh; and Nusrat Rahman, University of Dhak and Association for Advancement of Information Technology (AAIT), Bangladesh ........................................................................ 1591
Chapter 4.12 Chinese POS Disambiguation and Unknown Word Guessing with Lexicalized HMMs / Guohong Fu, The University of Hong Kong, Hong Kong; and Kang-Kwong Luke, The University of Hong Kong, Hong Kong ................................................................. 1595

Chapter 4.13 IT Implementation in a Developing Country Municipality: A Sociocognitive Analysis / Clive Sanford, Aalborg University, Denmark; and Anol Bhattacharjee, University of South Florida, USA ................................................................. 1608

Chapter 4.14 Personalization Issues for Science Museum Web Sites and E-Learning / Silvia Filippini-Fantoni, The University of Paris I Sorbonne University, France; Jonathan P. Bowen, London South Bank University, UK; Teresa Numerico, London South Bank University, UK ................................................................. 1629

Chapter 4.15 Personalization and Customer Satisfaction in Mobile Commerce / HY Sonya Hsu, Southern Illinois University, USA; and Songpol Kulviwat, Hofstra University, USA ................................................................. 1649

Chapter 4.16 Integrating Human Computer Interaction in Veterinary Medicine Curricula / Gale Parchoma, University of Saskatchewan, Canada; Susan M. Taylor, University of Saskatchewan, Canada; Jonathan M. Naylor, Ross University School of Veterinary Medicine, West Indies; Sameeh M. Abutarbush, University of Prince Edward Island, Canada; Katharina L. Lohmann, University of Saskatchewan, Canada; Kathy Schwarz, University of Saskatchewan, Canada; Cheryl Waldner, University of Saskatchewan, Canada; Sharon Porterfield, University of Saskatchewan, Canada; Cindy L. Shmon, University of Saskatchewan, Canada; Lydden Polley, University of Saskatchewan, Canada; Chris Clark, University of Saskatchewan Canada ................................................................. 1656

Chapter 4.17 Portals in Consumer Search Behavior and Product Customization / Ian Michael, Zayed University, UAE ................................................................. 1673

Chapter 4.18 Human Factors in the Development of Trend Detection and Tracking Techniques / Chaomei Chen, Drexel University, USA; Kaushal Toprani, Drexel University, USA; and Natasha Lobo, Drexel University, USA ................................................................. 1678

Chapter 4.19 The Internet, Health Information, and Managing Health: An Examination of Boomers and Seniors / Christopher G. Reddick, The University of Texas at San Antonio, USA ................................................................. 1687

Chapter 4.20 The Online Effect: Transitioning from the Legacy Help Desk to the Online Task Management System / Kym Mawson-Lee, University of New South Wales, Australia ................................. 1707

Section V. Organizational and Social Implications

This section includes a wide range of research pertaining to the social and organizational impact of human-computer interaction around the world. Chapters introducing this section analyze gender diff-
ferences in the adoption and reaction to IT, while later contributions offer an extensive analysis of differing social and cultural dimensions of technology adoption. The inquiries and methods presented in this section offer insight into the implications of human-computer interaction at both a personal and organizational level, while also emphasizing potential areas of study within the discipline.

Chapter 5.1 Gender, Race, Social Class and Information Technology / Myungsook Klassen, California Lutheran University, USA; and Russell Stockard, California Lutheran University, USA................................................................. 1729

Chapter 5.2 Gender and the Culture of Computing in Applied IT Education / Susan C. Herring, Indiana University, USA; Christine Ogan, Indiana University, USA; Manju Ahuja, Indiana University, USA; and Jean C. Robinson, Indiana University, USA.......... 1736

Chapter 5.3 Gender Equalization in Computer-Mediated Communication / Rosalie J. Ocker, The Pennsylvania State University, USA ................................................................. 1745

Chapter 5.4 The Cross-Cultural Dimension of Gender and Information Technology / Haiyan Huang, The Pennsylvania State University, USA........................................................................ 1753

Chapter 5.5 Enhancing E-Collaboration Through Culturally Appropriate User Interfaces / Dianne Cyr, Simon Fraser University, Canada ...................................................... 1761

Chapter 5.6 Cultural Barriers of Human-Computer Interaction / Deborah Sater Carstens, Florida Institute of Technology, USA ........................................................................ 1769

Chapter 5.7 Technology and Culture: Indian Experiences / Ramesh C. Sharma, Indira Gandhi National Open University, India; and Sanjaya Mishra, Indira Gandhi National Open University, India...................................................... 1777

Chapter 5.8 Intercultural Computer-Mediated Communication Between Chinese and U.S. College Students / Yun Xia, Rider University, USA ............................................................. 1786

Chapter 5.9 Global Culture and Computer Mediated Communication / Susan R. Fussell, Carnegie Mellon University, USA; Qiping Zhang, Long Island University, USA; and Leslie D. Setlock, Carnegie Mellon University, USA................................................................. 1801

Chapter 5.10 Linguistics of Computer-Mediated Communication: Approaching the Metaphor / Rosanna Tarsiero, Gionnethics, Italy .............................................................. 1817

Chapter 5.11 Impression Formation in Computer-Mediated Communication and Making a Good (Virtual) Impression / Jamie S. Switzer, Colorado State University, USA .... 1837

Chapter 5.12 Group Decision Making in Computer-Mediated Communication as Networked Communication: Understanding the Technology and Implications / Bolanle A. Olaniran, Texas Tech University, USA ........................................................................ 1849

Chapter 5.13 Effects of Computer-Mediated Communication / Stuart S. Gold, DeVry University, USA ........................................................................ 1864
Chapter 5.14 Exploring “Events” as an Information Systems Research Methodology / Anita Greenhill, The University of Manchester, UK; and Gordon Fletcher, University of Salford, UK ................................................................. 1871

Chapter 5.15 Culture’s Impact on Technology Mediated Learning: The Role of Horizontal and Vertical Individualism and Collectivism / Steven Hornik, University of Central Florida, USA; and Anna Tupchiy, University of Central Florida, USA ........................................................................................................ 1885

Chapter 5.16 Using Social Networking Analysis to Facilitate Knowledge Sharing in the British Council / Bonnie Wai-yi Cheuk, Improvement Service for the Scottish Local Government, Scotland ........................................................................................................ 1905

Chapter 5.17 Collaboration Challenges in Community Telecommunication Networks / Sylvie Albert, Laurentian University, Canada; Rolland LeBrasseur, Laurentian University, Canada ........................................................................................................ 1915

Chapter 5.18 Planning and Managing the Human Factors for the Adoption and Diffusion of Object-Oriented Software Development Processes / Magdy K. Serour, University of Technology, Sydney, Australia ........................................................................................................ 1937

Chapter 5.19 Product Customization on the Web: An Empirical Study of Factors Impacting Choiceboard User Satisfaction / Pratyush Bharati, University of Massachusetts, Boston, USA; and Abhijit Chaudhury, Bryant University, USA ........................................................................................................ 1956

Chapter 5.20 Distributed Deception: An Investigation of the Effectiveness of Deceptive Communication in a Computer-Mediated Environment / Randall J. Boyle, University of Utah, USA; Charles J. Kacmar, The University of Alabama, USA; and Joey F. George, Florida State University, USA ........................................................................................................ 1970

Chapter 5.21 Extrinsic Plus Intrinsic Human Factors Influencing the Web Usage / Manuel Jesús Sánchez-Franco, University of Seville, Spain ........................................................................................................ 1995

Section VI. Managerial Impact

This section presents contemporary coverage of the managerial implications of human-computer interaction. Particular contributions address virtual networking, knowledge blogs in company environments, and ambient business. The managerial research provided in this section allows executives, practitioners, and researchers to gain a better sense of the relationship between of business, technology, and individuals.

Chapter 6.1 Social Impact of Virtual Networking / Hakikur Rahman, SDNP Bangladesh, Bangladesh ........................................................................................................ 2014

Chapter 6.2 Gender and Telework in Information Technology / Paula F. Saddler, Association of Certified Fraud Examiners, USA; Donald D. Davis, Old Dominion University, USA;
Chapter 6.3 Managerial Careers, Gender, and Information Technology Field / Iiris Aaltio, Lappeenranta University of Technology, Finland ................................................................. 2030

Chapter 6.4 Social Networking and Knowledge Transfer in Collaborative Product Development / Katariina Ala-Rämi, University of Oulu, Finland .............................................................................. 2037

Chapter 6.5 Reframing Information System Design as Learning Across Communities of Practice / Kevin Gallagher, Northern Kentucky University, USA; and Robert M. Mason, University of Washington, USA ........................................................................ 2052

Chapter 6.6 Grounding Business Interaction Models: Socio-Instrumental Pragmatism as a Theoretical Foundation / Göran Goldkuhl, Linköping University & Jönköping International Business School, Sweden; and Mikael Lind, University College of Borås, Sweden, & Linköping University, Sweden ................................................................. 2071

Chapter 6.7 Managing Socio-Technical Integration in Iterative Information System Development Projects / Bendik Bygstad, Norwegian School of Information Technology, Norway. 2090

Chapter 6.8 Human Factors for Networked and Virtual Organizations / Vincent E. Lasnik, Independent Knowledge Architect, USA .................................................................................. 2106

Chapter 6.9 Trust in Computer-Mediated Communications: Implications for Individuals and Organizations / Susan K. Lippert, Drexel University, USA ..................................................................... 2118

Chapter 6.10 Understanding an ERP System Implementation in a Higher Education Institution: A Grounded Theory Approach / Jose Esteves, Universitat Politècnica de Catalunya, Barcelona, Spain; and Joan Pastor, Universitat Internacional de Catalunya, Barcelona, Spain ........................................................................................................ 2132

Chapter 6.11 Ontologies for Scalable Services-Based Ubiquitous Computing / Daniel Oberle, SAP Research, CEC Karlsruhe, Germany; Christof Bornhövd, SAP Research, Research Center Palo Alto, USA; and Michael Altenhofen, SAP Research, CEC Karlsruhe, Germany .................................................................................. 2144

Chapter 6.12 Web Personalization for E-Marketing Intelligence / Penelope Markellou, University of Patras, Greece & Research Academic Computer Technology Institute, Greece; Maria Rigou, University of Patras, Greece & Research Academic Computer Technology Institute, Greece; and Spiros Sirmakessis, Technological Educational Institution of Messolongi, Greece & Research Academic Computer Technology Institute, Greece ........................................................................ 2164

Chapter 6.13 Knowledge Blogs in Firm Internal Use / Miia Kosonen, Lappeenranta University of Technology, Finland;
Kaisa Henttonen, Lappeenranta University of Technology, Finland; and Kirsimarja Blomqvist, Lappeenranta University of Technology, Finland ........................................ 2181

Chapter 6.14 Towards Ambient Business: Enabling Open Innovation in a World of Ubiquitous Computing / Christian Schmitt, University of Cologne, Germany; Detlef Schoder, University of Cologne, Germany; Kai Fischbach, University of Cologne, Germany; and Steffen Muhle, University of Cologne, Germany ................................................................. 2191

Chapter 6.15 Successes and Failures of SAP Implementation: A Learning Perspective / Tanya Bondarouk, University of Twente, The Netherlands; and Maarten van Riemsdijk, University of Twente, The Netherlands ................................................................................................................. 2214

Chapter 6.16 The Human Face of E-Business: Engendering Consumer Initial Trust Through the Use of Images of Sales Personnel on E-Commerce Web Sites / Khalid Aldiri, University of Bradford, UK; Dave Hobbs, University of Bradford, UK; Rami Qahwaji, University of Bradford, UK ................................................................................................................. 2235

Section VII. Critical Issues

This section addresses conceptual and theoretical issues related to the field of human-computer interaction. Within these chapters, the reader is presented with analysis of the most current and relevant conceptual inquires within this growing field of study. Particular chapters address particular algorithms for web personalization, sociotechnical change, and security for ubiquitous computing systems. Overall, contributions within this section ask unique, often theoretical questions related to the study of human-computer interaction and, more often than not, conclude that solutions are both numerous and contradictory.

Chapter 7.1 Gender, Place, and Information Technology / Anita Greenhill, The University of Manchester, UK ................................................................................................................. 2259

Chapter 7.2 The Intersection of Gender, Information Technology, and Art / Linda Doyle, University of Dublin, Ireland; and Maryann Valiulis, University of Dublin, Ireland ................................................................................................................. 2266

Chapter 7.3 Gender-Based Attitudes Toward Technology / Konrad Morgan, University of Bergen, Norway; and Madeleine Morgan, University of Bergen, Norway ................................................................................................................. 2274

Chapter 7.4 Gender in Distance Education Technology / Colette Wanless-Sobel, University of Minnesota & Inver Hills Community College, USA ................................................................................................................. 2278

Chapter 7.5 Gender Differences in Information Technology Acceptance / Will Wai-kit Ma, Hong Kong University of Science and Technology, China and Hong Kong Shue Yan College, China; and Allan Hio-kau Yuen, The University of Hong Kong, China ................................................................................................................. 2287

Chapter 7.6 Discourses in Gender and Technology: Taking a Feminist Gaze / Sheila French, Manchester Metropolitan University, UK ................................................................................................................. 2296
Chapter 7.7 Theorizing Gender and Information Technology Research /
Eileen M. Trauth, The Pennsylvania State University, USA ................................................................. 2309

Chapter 7.8 Socio-Cognitive Model of Trust /
Rino Falcone, Institute for Cognitive Sciences and Technology, National Research Council of Italy, Italy; and Cristiano Castelfranchi, Institute for Cognitive Sciences and Technology National Research Council of Italy, Italy ............................................................................................................ 2316

Chapter 7.9 Information-Communications Systems Convergence Paradigm:
Invisible E-Culture and E-Technologies / Fjodor Ruzic, Institute for Informatics, Croatia .............. 2322

Chapter 7.10 IT Artefacts as Socio-Pragmatic Instruments: Reconciling the Pragmatic,
Semiotic, and Technical / Göran Goldkuhl, Linköping University &
Jönköping International Business School, Sweden;
and Pär J. Ågerfalk, University of Limerick, Ireland & Örebro University, Sweden ..................... 2341

Chapter 7.11 Ignorance is Bliss: The Effect of Increased Knowledge
on Privacy Concerns and Internet Shopping Site Personalization Preferences /
Thomas P. Van Dyke, The University of North Carolina Greensboro, USA ........................................ 2356

Chapter 7.12 The Function of Representation in a “Smart Home Context” /
Mats Edenius, Stockholm School of Economics, Sweden ............................................................................ 2376

Chapter 7.13 Towards a Meta-Model for Socio-Instrumental Pragmatism /
Peter Rittgen, University College of Borås, Sweden ........................................................................... 2390

Chapter 7.14 The Effect of Choice and Announcement Duration on
the Estimation of Telephone Hold Time / Philip Kortum, Rice University, USA;
Randolph G. Bias, The University of Texas at Austin, USA;
Benjamin A. Knott, U.S. Air Force Research Laboratory, USA;
and Robert G. Bushey, AT&T Laboratories, USA ................................................................................... 2404

Chapter 7.15 Towards Adaptive Security for Ubiquitous Computing Systems:
MOSQUITO and Serenity / Volkmar Lotz, SAP Research, France;
Luca Compagna, SAP Research, France; and Konrad Wrona, SAP Research, France ...................... 2424

Chapter 7.16 Trusting Computers Through Trusting Humans: Software Verification
in a Safety-Critical Information System / Alison Adam, University of Salford, UK;
and Paul Spedding, University of Salford, UK ...................................................................................... 2436

Chapter 7.17 Global Information Ethics: The Importance of Being Environmentally Earnest /
Luciano Floridi, Universita degli Studi di Bari, Italy and Oxford University, UK ......................... 2450
Chapter 7.18 Emotional Digitalization as Technology of the Postmodern: 
A Reflexive Examination from the View of the Industry / 
Claus Hohmann, Bergische Universitaet Wuppertal, Germany .................................................. 2462

Chapter 7.19 The Paleolithic Stone Age Effect? Gender Differences Performing 
Specific Computer-Generated Spatial Tasks / 
Geoffrey S. Hubona, Georgia State University, USA; 
and Gregory W. Shirah, National Aeronautics and Space Administration, USA ......................... 2473

Chapter 7.20 Optimality-Theoretic Lexical Mapping Theory: A Case Study of 
Locative Inversion / One-Soon Her, National Chengchi University, Taiwan ............................... 2499

Section VIII. Emerging Trends

This section highlights research potential within the field of human-computer interaction while exploring 
uncharted areas of study for the advancement of the discipline. Chapters within this section highlight 
the socio-semantic web, podcasts, and the emergence of increasingly intelligent user interfaces. These 
contributions, which conclude this exhaustive, multi-volume set, provide emerging trends and sugges-
tions for future research within this rapidly expanding discipline.

Chapter 8.1 Podcastia: Imagining Communities of Pod-People / Jonathan Cohn, UCLA, USA .... 2530

Chapter 8.2 Maximizing Web Accessibility Through User-Centered Interface Design / 
Soonhwa Seok, The University of Kansas, USA ............................................................................ 2542

Chapter 8.3 Voice-Enabled User Interfaces for Mobile Devices / 
Louise E. Moser, University of California, Santa Barbara, USA; 
and P. M. Melliar-Smith, University of California, Santa Barbara, USA ........................................ 2556

Chapter 8.4 Multilayered Approach to Evaluate Mobile User Interfaces / 
Maria de Fátima Queiroz Vieira Turnell, Universidade Federal de Campina Grande (UFCG), Brazil; 
José Eustáquio Rangel de Queiroz, Universidade Federal de Campina Grande (UFCG), Brazil; 
Danilo de Sousa Ferreira, Universidade Federal de Campina Grande (UFCG), Brazil .............. 2573

Chapter 8.5 Digital Imaging Trek: A Practical Model for Managing 
the Demand of the Digitally Enabled Traveller / Stephen C. Andrade, Johnson 
& Wales University, USA; and Hilary Mason, Johnson & Wales University, USA ................... 2590

Chapter 8.6 Intelligent User Interfaces for Mobile Computing / 
Michael J. O’Grady, University College Dublin, Ireland; 
and Gregory M. P. O’Hare, University College Dublin, Ireland ............................................. 2614

Chapter 8.7 Intelligent User Interfaces for Ubiquitous Computing / 
Rainer Malaka, Bremen University, Bremen .................................................................................. 2627
Chapter 8.8 Vive la Différence: The Cross-Culture Differences Within Us /
David Gefen, Drexel University, USA; Nitza Geri, The Open University of Israel, Israel;
and Narasimha Paravastu, Central State University, USA ...................................................... 2644

Chapter 8.9 Growing Up Wireless: Being a Parent and Being a Child in
the Age of Mobile Communication / Letizia Caronia, University of Bologna, Italy ................. 2655

Chapter 8.10 Politeness as a Social Computing Requirement /
Brian Whitworth, Massey University, New Zealand;
and Tong Liu, Massey University, New Zealand ........................................................................... 2675

Chapter 8.11 Neo-Symbiosis: The Next Stage in the Evolution
of Human Information Interaction /
Douglas Griffith, General Dynamics Advanced Information Systems, USA;
and Frank L. Greitzer, Pacific Northwest National Laboratory, USA ........................................ 2693

Chapter 8.12 Enabling Programmable Ubiquitous Computing Environments:
A Middleware Perspective / Christine Julien, The University of Texas at Austin, USA;
and Sanem Kabadayi, The University of Texas at Austin, USA ...................................................... 2708
Preface

The use of, engagement with, and study of technology has become the basis of modern life. As such, the field of human-computer interaction has emerged as an essential, multidisciplinary field that seeks to determine how humans interact with different interfaces and how these interfaces can be better designed, evaluated, and implemented to minimize barriers between human thought and computer behavior. As we continue to develop new and innovative technologies, we must also strive to more fully understand how these technologies impact humanity, both on a societal and individual level, and to remain aware of the latest in human-computer interaction research and exploration.

In recent years, the areas of study related to the interaction between people and technology have become innumerable. As a result, researchers, practitioners, and educators have devised a variety of techniques and methodologies to develop, deliver, and, at the same time, evaluate the effectiveness of the interfaces implemented and used in modern society. The explosion of methodologies in the field has created an abundance of new, state-of-the-art literature related to all aspects of this expanding discipline. This body of work allows researchers to learn about the fundamental theories, latest discoveries, and forthcoming trends in the field of human-computer interaction.

Constant technological and theoretical innovation challenges researchers to remain informed of and continue to develop and deliver methodologies and techniques utilizing the discipline’s latest advancements. In order to provide the most comprehensive, in-depth, and current coverage of all related topics and their applications, as well as to offer a single reference source on all conceptual, methodological, technical, and managerial issues in human-computer interaction, Information Science Reference is pleased to offer a four-volume reference collection on this rapidly growing discipline. This collection aims to empower researchers, practitioners, and students by facilitating their comprehensive understanding of the most critical areas within this field of study.

This collection, entitled Human Computer Interaction: Concepts, Methodologies, Tools, and Applications, is organized into eight distinct sections which are as follows: 1) Fundamental Concepts and Theories, 2) Development and Design Methodologies, 3) Tools and Technologies, 4) Utilization and Application, 5) Organizational and Social Implications, 6) Managerial Impact, 7) Critical Issues, and 8) Emerging Trends. The following paragraphs provide a summary of what is covered in each section of this multi-volume reference collection.

Section One, Fundamental Concepts and Theories, serves as a foundation for this exhaustive reference tool by addressing crucial theories essential to understanding human-computer interaction. Specific issues in human-computer interaction, such as ubiquitous computing, communities of practice, and online social networking are discussed in selections such as “Ubiquitous Computing and the Concept of Context” by Antti Oulasvirta and Antti Salovaara, “Sociotechnical Theory and Communities of Practice” by Andrew Wenn, and “Online Social Networking for New Researched Opportunities” by Lionel Mew. Within the selection “Personalization Techniques and Their Application,” authors Juergen Anke
and David Sundaram provide an overview of the personalization of information systems and the impact such an approach has on the usability of everything from e-learning environments to mobile devices. The selections within this comprehensive, foundational section allow readers to learn from expert research on the elemental theories underscoring human-computer interaction.

Section Two, Development and Design Methodologies, contains in-depth coverage of conceptual architectures and frameworks, providing the reader with a comprehensive understanding of emerging theoretical and conceptual developments within the development and utilization of tools and environments that promote interaction between people and technology. Beginning this section is the contribution “Measuring the Human Element in Complex Technologies” by Niamh McNamara and Jurek Kirakowski, which analyzes the role that user satisfaction has in shaping the design of products and the impact this has on developers. Similarly, “Design Methods for Experience Design” by Marie Jefsioutine and John Knight describes the authors’ framework for web site design—the Experience Design Framework. Other selections, such as “Content Personalization for Mobile Interfaces” by Spiridoula Koukia, Maria Rigou, and Spiros Sirmakessis and “Kinetic User Interfaces: Physical Embodied Interaction with Mobile Ubiquitous Computing Systems” by Vincenzo Pallotta, Pascal Bruegger, and Béat Hirshbrunner offer insight into particular methodologies for the design, development, and personalization of mobile user interfaces. From basic designs to abstract development, chapters such as “Designing and Evaluating In-Car User-Interfaces” by Gary Burnett and “Integrating Usability, Semiotic, and Software Engineering into a Method for Evaluating User Interfaces” by Kenia Sousa, Albert Schilling, and Elizabeth Furtado serve to expand the reaches of development and design techniques within the field of human-computer interaction.

Section Three, Tools and Technologies, presents extensive coverage of various tools and technologies that individuals interact, collaborate, and engage with every day. The emergence of social networking tools such as blogs and wikis is explored at length in selections such as “PDA Plagiarism, Instruction, and Blogs” by Michael Hanrahan, “Wikis as Tools for Collaboration” by Jane Klobas, and “Assessing Weblogs as Education Portals” by Ian Weber. Further discussions of the role of technology in learning are explored in chapters such as “How Technology Can Support Culture and Learning” by David Luigi Fuschi, Bee Ong, and David Crombie and “Facilitating E-Learning with Social Software: Attitudes and Usage from the Student’s Point of View” by Reinhard Bernsteiner, Herwig Ostermann, and Roland Staudinger. The latter of these two chapters discusses both the theoretical basis for the implementation of technology in an educational setting and the impact this implementation has on students. Ultimately, the authors conclude that online social networking has the potential to emerge as a useful tool for both students and teachers. The rigorously researched chapters contained within this section offer readers countless examples of the dynamic interaction between humans and technology.

Section Four, Utilization and Application, provides an in-depth analysis of the practical side of human computer interaction, focusing specifically on environments in which the relationship between humans and technology has been significant. Mobile device usage is highlighted in the selections “A De-Construction of Wireless Device Usage” by Mary R. Lind, “Localized User Interface for Improving Cell phone Users’ Device Competency” by Lucia D. Krisnawati and Restyandito, and ”Mobile Phone Use Across Cultures: A Comparison Between the United Kingdom and Sudan” by Ishraga Khattab and Steve Love. These selections offer conceptualization and analysis of the factors impacting wireless device usage, ultimately determining that many factors, such as culture, familiarity with the device itself, and ease of use influence an individual’s wireless device usage. Further contributions explore how human factors impact the successful use of technology in areas such as trend detection, veterinary medicine, mobile commerce, and web browsing. From established applications to forthcoming innovations, contributions in this section provide excellent coverage of today’s global community and demonstrate
how the interaction between humans and technology impacts the social, economic, and political fabric of our present-day global village.

Section Five, **Organizational and Social Implications**, includes a wide range of research pertaining to the organizational and cultural implications of humans’ reaction to technology. This section begins with a thorough analysis of the intersection of gender and technology in contributions such as “Gender, Race, Social Class, and Information Technology” by Myungsook Klassen and Russell Stockard, Jr., “Gender and the Culture of Computing in Applied IT Education” by Susan C. Herring, Christine Ogan, Manju Ahuja, and Jean C. Robinson, and “Gender Equalization in Computer-Mediated Communication” by Rosalie J. Ocker. Other issues that are surveyed within this section include the implication of cultural differences within Deborah Sater Carstens’ “Cultural Barriers of Human-Computer Interaction,” computer mediated communication in Bolanle A. Olaniran’s “Group Decision Making in Computer-Mediated Communication as Networked Communication: Understanding the Technology and Implications,” and community telecommunication networks in Sylvie Albert and Rolland LeBrasseur’s “Collaboration Challenges in Community Telecommunication Networks.” Overall, the discussions presented in this section offer insight into the implications of human computer interaction in both organizational and social settings, as well as provide solutions for existing problems and shortcomings.

Section Six, **Managerial Impact**, presents contemporary coverage of the managerial applications and implications of human computer interaction. This collection of research opens with “Social Impact of Virtual Networking” by Hakikur Rahman, which documents the emergence of the virtual enterprise and the impact such a structure has on social communication. Similarly, within the article “Human Factors for Networked and Virtual Organizations,” Vincent E. Lasnik emphasizes the role that human factors engineering must play in the future design of virtual and networked environments. Later contributions, such as “Knowledge Blogs in Firm Internal Use,” investigate knowledge transfer within and among organizations. Within this selection, authors Miia Kosonen, Kaisa Henttonen, and Kirsimarja Blomqvist identify factors for the implementation of “knowledge blogs” within organizations and explain why these tools can be used to encourage institutional memory, create identities, and to inform the organization itself. The comprehensive research in this section offers an overview of the major issues that practitioners, managers, end users and even consumers must address in order to remain informed about the latest managerial changes in the field of human computer interaction.

Section Seven, **Critical Issues**, presents readers with an in-depth analysis of the more theoretical and conceptual issues within this growing field of study by addressing topics such as security, ethics, and gender differences in technology adoption and use. Specifically, these topics are discussed in selections such as “Trusting Computers Through Trusting Humans: Software Verification in a Safety-Critical Information System” by Alison Adam and Paul Spedding and “Global Information Ethics: The Importance of Being Environmentally Earnest” by Luciano Floridi. Later selections, which include “Emotional Digitalization as Technology of the Postmodern: A Reflexive Examination from the View of the Industry,” review more novel issues, such as how the digitalization of emotions helps to bridge the gap between technology and humanity. Specifically, within this chapter, author Claus Hohmann identifies the emotion inherent in the design of new technologies and investigates how this has and will continue to impact human reaction. In all, the theoretical and abstract issues presented and analyzed within this collection form the backbone of revolutionary research in and evaluation of human-computer interaction.

The concluding section of this authoritative reference tool, **Emerging Trends**, highlights research potential within the field of human-computer interaction while exploring uncharted areas of study for the advancement of the discipline. “Communicating in the Information Society: New Tools for New Practices” by Lorenzo Cantoni and Stefano Tardini presents a framework for the latest digital communication tools and their implementation, while the evolution of the semantic web is analyzed in Cristian Peraboni
and Laura A. Ripamonti’s “Socio-Semantic Web for Sharing Knowledge.” The nature of podcasts and their role in encouraging collaboration and acting as an educational tool is explored in “Stay Tuned for Podcast U and the Data on M-Learning” by Deborah Vess and Michael Gass and “Podcastia: Imagining Communities of Pod-People by Jonathan Cohn. Other new trends, such as voice enabled interfaces for mobile devices, programmable ubiquitous computing environments, and intelligent user interfaces for mobile and ubiquitous computing are discussed in this collection. This final section demonstrates that humanity’s interaction with technology will continue to grow and evolve, shaping every facet of modern life.

Although the contents of this multi-volume book are organized within the preceding eight sections which offer a progression of coverage of important concepts, methodologies, technologies, applications, social issues, and emerging trends, the reader can also identify specific contents by utilizing the extensive indexing system listed at the end of each volume. Furthermore, to ensure that the scholar, researcher, and educator have access to the entire contents of this multi-volume set, as well as additional coverage that could not be included in the print version of this publication, the publisher will provide unlimited, multi-user electronic access to the online aggregated database of this collection for the life of the edition free of charge when a library purchases a print copy. In addition to providing content not included within the print version, this aggregated database is also continually updated to ensure that the most current research is available to those interested in human-computer interaction.

As technology continues its rapid advancement, the study of how to successfully design, implement and, ultimately, evaluate human reactions to and interactions with the modern world becomes increasingly critical. Innovations in the design of mobile devices, educational environments, and web sites have all been made possible through a more thorough understanding of how humans react to and engage with technological interfaces. Continued evolution in our understanding of the human-computer dynamic will encourage the development of more usable interfaces and models that aim to more thoroughly understand the theories of successful human-computer interaction.

The diverse and comprehensive coverage of human-computer interaction in this four-volume, authoritative publication will contribute to a better understanding of all topics, research, and discoveries in this developing, significant field of study. Furthermore, the contributions included in this multi-volume collection series will be instrumental in the expansion of the body of knowledge in this enormous field, resulting in a greater understanding of the fundamentals while also fueling the research initiatives in emerging fields. We at Information Science Reference, along with the editor of this collection, hope that this multi-volume collection will become instrumental in the expansion of the discipline and will promote the continued growth of human-computer interaction.
FUNDAMENTAL CONCEPTS AND THEORIES IN HUMAN COMPUTER INTERACTION

User-centered design begins with a thorough understanding of the needs and requirements of the users (Shneiderman, 1992). User needs analysis (or Knowledge Elicitation) and evaluation methods in HCI are a critical process to the success of requirements and design gathering (Maiden et al., 1995), usability testing, and user evaluation stages of software development (Zaphiris & Kurniawan, 2001).

Examples of knowledge elicitation methods often start with initial questionnaire feedback, requirements task walkthroughs, interviews techniques, and focus group sessions. They can rapidly scale upwards to more complex psychometric design and evaluation processes. For example, various fidelities of prototype construction, direct and indirect observation practices for monitoring user actions and response time comparisons, and methods for eliciting mental categorization models, that is, in distinguishing expert and non-expert technology usage patterns.

The measure of a good experience can vary from person to person; however the appropriate understanding of a usable design comes from gaining the knowledge that it is functional, efficient, and desirable to its intended audience (Kuniavksy, 2003).

John and Marks (1997) identify three key factors to assess the usability of an interface:

Usability is measured by the extent to which the intended goals of use of the overall system are achieved (effectiveness), the resources that have to be expended to achieve the intended goals (efficiency), and the extent to which the user finds the overall system acceptable (satisfaction) (John and Marks, 1997).

The usability of a system is also related to issues surrounding its accessibility. There is a broad range of users to whom Web-based services are directed, and the services provided ought to be accessible to them (e.g. visually, hearing, physically or cognitively impaired or even people with different experience of and attitudes towards technology).
In the United Kingdom, “The Disability Discrimination Act” (DDA) began to come into effect in December 1996 and brought in measures to prevent discrimination against people on the basis of disability. Part III of the Act aims to ensure that disabled people have equal access to products and services. Under Part III of the Act, businesses that provide goods, facilities, and services to the general public (whether paid for or free) need to make reasonable adjustments for disabled people to ensure they do not discriminate by:

- Refusing to provide a service;
- providing a service of a lower standard or in a worse manner; and
- providing a service on less favorable terms than they would to users without the disability.

There is a legal obligation on service providers to ensure that disabled people have equal access to Web-based products and services. Section 19(1) (c) of the Act makes it unlawful for a service provider to discriminate against a disabled person “in the standard of service which it provides to the disabled person or the manner in which it provides it”.

An important provision here is that education is not covered by the DDA, but by separate legislation, titled the Special Educational Needs and Disability Act 2001 (SENDA). This Act introduces the right for disabled students not to be discriminated against in education, training, and any services provided wholly or mainly for students, and for those enrolled in courses provided by “responsible bodies”, including further and higher education institutions and sixth form colleges. Student services covered by the Act can include a wide range of educational and non-educational services, such as field trips, examinations and assessments, short courses, arrangements for work placements and libraries, and learning resources. In a similar wording to the DDA, SENDA requires responsible bodies to make reasonable adjustments so that people with disabilities are not at a substantial disadvantage.

DEVELOPMENT AND DESIGN METHODOLOGIES

There are many different elicitation and usability/accessibility evaluation techniques (Cooke, 1994) and selecting the “right” technique in a particular situation is not trivial.

Burge’s Table of Knowledge Elicitation Methods (Burge, 2001) provides an extensive comparative view of almost all the common KE techniques found in HCI.

In addition, usability and accessibility evaluation techniques are often grouped into two broad areas: user-based (that often include user testing) and expert based (that often include heuristic evaluation and cognitive walkthrough) techniques.

King et al. (2004) presented what they called “An incremental usability and accessibility evaluation framework for digital libraries”. Their framework is broken down into seven key activities and addresses all stages of a design of a system.

**Activity 1: Conduct Query - Requirement Gathering**
Identify satisfaction levels of current users of the system and establish key positive and negative aspects of the interface, what features they would like to see, and so forth.

**Activity 2: Analysis**
Evaluate current findings and identify issues not yet addressed
Activity 3: Perform Empirical (user) Evaluations
We regard user testing as the strongest evaluation technique, allowing us to identify real user problems by observing users interacting with the system. Retrospective focus groups or interviews conducted after the evaluations also provide a volume of qualitative data.

Activity 4: Analysis
Establish key problems and assess if any areas of the service have not been covered by user evaluations

Activity 5: Expert Evaluations
Appropriate modification of expert evaluation techniques maybe required so that they supplement previous evaluation findings, and address any areas or issues that have not as yet been covered

Activity 6: Analysis
Analyze all data identifying key issues that need to be addressed in the redesign of the service.
Establish new usability and accessibility goals for the design

Activity 7: Iterative Process
Re-conduct all stages in the iterative framework to evaluate redesign

We now describe some of the key methods that are associated with the above mentioned framework.

Interviewing

This query-based process elicits from users knowledge on a set information topic based on their expertise in the domain in question. It is useful for obtaining behavioral reasoning and background knowledge. Interviews can be categorized as structured or unstructured. Structured interviews elicit limited responses from users, by using a series of closed questions that have to be answered based on given solutions. This enables the user data to be analysed quicker but is not necessarily as informative as unstructured (open ended) interviews.

Preece et al. (1994) suggests that interview processes are most effective as semi-structured based on a series of fixed questions that gradually lead into more in-depth user needs and requirements understanding, then allowing for open ended responses to possibly create new dynamic questions based on prior structured responses (Macaulay, 1996). On-site stakeholder interviews allows researchers to bring about a vivid mental model of how users work with existing systems and how new systems can support them (Mander and Smith, 2002).

Interviews are useful when combined with surveys or questionnaires, as they can be used to improve the validity of data by clarifying specific issues that were raised in the survey or questionnaire.

Surveys

In conducting surveys, three things are necessary: a) the set of questions, b) a way to collect responses, and c) access to the demographics group you wish to test. There are several widely reported templates for acquiring different types of user data, such as the Quality of User Interface Satisfaction (QUIS) by Chin et al. (1988) and the Computer System Usability Questionnaire (CSUQ) by IBM with Lewis et al. (1995).
Surveys can be similarly open and closed question based, but also allow us to enquire scalar results giving indicators of quality in positive and negative statements. Self-filling surveys can be time efficient to deploy, and results from closed questions can be fast to analyze.

Open questions tend to elicit unanticipated information which can be very useful for early design. Existing survey sampling techniques include face-to-face, paper- and pencil-based, telephone surveys where the researcher will fill in the results (which becomes more of an interview style) but there is modern interest in computer assisted and Web-based surveying techniques.

Focus Groups

This activity is useful for eliciting cross-representative domains of knowledge from several stakeholders/users, in an open discussion format. Sessions are often moderated and tend to be informal by nature, centering on the creation of new topics from open questions.

Evidence shows that the optimal number needed for a mixed experience focus group is between five to eight participants, with group size being inversely related to the degree of participation (Millward, 1995).

Observation

Observation methods elicit user knowledge from the way users interact with a prototype or a final product. It can be direct, whereby a researcher is present and can steer users to particular points in an interaction. This tends to utilize video camera equipment and note taking to successfully enquire the timeline of user actions that is “getting from point A to point D may require steps B or C”.

The other model of observation is indirect, whereby all user actions are captured electronically. The researcher has to maintain co-operation between users and should only pose questions if clarification is needed.

Paper Prototyping

There are several approaches to paper prototypes, enabling users to create quick and partial designs of their concepts. It is often used in early stages of the design processes. Though the methodology lacks standardization, Rettig (1994) distinguishes between high-tech and low-tech views, and the more commonly modeled categories are of low, medium and high fidelity prototypes (Greenberg, 1998). Rudd et al. (1996) also distinguishes prototypes according to horizontal and vertical prototypes, with vertical representing deep functionality of a limited view to the final output, and horizontal giving a wide overview of the full functionality of the system but with a weaker depth of understanding. Hall (2001) discusses the benefits of using various levels of fidelities of prototypes.

Cognitive Walkthrough

Cognitive Walkthrough is an expert based evaluation technique that steps through a scenario/task by focusing on the users’ knowledge and goals. The expert evaluator first starts with descriptions of: the prototype interface, the task(s) from the user’s perspective, the correct sequence of actions needed to complete the task using the prototype and any assumptions about the characteristics of the user.

Then the evaluator walks through the tasks using the system, reviewing the actions that are necessary and attempting to predict how the users will behave.
A series of key questions are used throughout each sub-task evaluation:

- Will the user be trying to achieve the correct effect?
- Will the user know that the correct action is available?
- Will user know that the correct action will achieve the desired effect?
- If the correct action is taken, will the user see that things are going okay?

**Heuristic Evaluation**

Heuristic Evaluation is an expert review technique where experts inspect the interface to judge compliance with established usability principles (the ‘heuristics’)

Heuristic Evaluation is usually conducted in a series of four steps:

- Prepare: create a prototype to evaluate; select evaluators; prepare coding sheets to record problems
- Determine approach: either set typical user tasks (probably the most useful approach) or allow evaluators to establish their own tasks or conduct an exhaustive inspection of entire interface
- Conduct the evaluation: evaluators inspect interface individually to identify all violations of heuristics (the usability problems); record the problem (feature and location), severity (based on frequency, impact, criticality/cost) and heuristic violated
- Aggregate and analyze results: group similar problems; reassess severity; determine possible fixes

**Hierarchical Task Analysis (HTA)**

Hierarchical task analysis (HTA) is an analytic tool commonly used to describe people’s actions when interacting with the artifacts. We can analyze people’s work or leisure activities by focusing on what people do, their goals, mediating artifacts and their background knowledge. The primary use of HTA is to understand, clarify and organize our knowledge about existing system and work. It is applied in design and evaluation of the interactive system.

Essentially, HTA is concerned with the structure of tasks people are engaged in and how the tasks can be broken down into steps of sub-task hierarchies and how the order of sub-tasks can be described meaningfully.

In HTA, we break down a task from top to bottom, showing a hierarchical relationship amongst the tasks. This allows us to view the task at different levels of details

A task is the set of activities in which a user engages to achieve a goal. In other words, a task consists of a sequence of actions performed to achieve a goal. A goal is defined as the desired state of a system.

Usually HTA involves collecting information using data collection techniques (e.g. interviews, questionnaires, observations) and then analyzing it.

HTA can be presented as text or diagrams to show the hierarchy.

There are generally eight steps for task analysis:

- Define the purpose of analysis and the boundary of the system.
- Identify user groups, select representatives, and identify main tasks of concern.
- Design and conduct data collection to elicit information about these tasks.
- Identify the goals (and sub-goals) that users are trying to achieve, the steps they engage in, their rationale, and the information resources they use.
• Obtain information from documentation, interviews, questionnaires, focus groups, observation, ethnography, experiments, and so forth.
• Analyze the data to model specific instances of tasks initially, and then consider decomposition of tasks, balance of the model and stopping rules.
• Generalize across the specific task models to create a generic (or composite) task model: From each task model for the same goal produce a generic model that includes all the different ways of achieving the goal.
• Check the models with users, other stakeholders, analysts, and iterate the analysis process.

TOOLS AND TECHNOLOGIES

HCI community approaches design problems using different tools to help identify the users, their needs and behaviors, user interface problems. Various kinds of tools have been developed to tackle different research objectives. We can categorize these tools into three major categories: hardware, software and analytic tools. In this section we will describe an example of tool for each category

Eye Tracking

Eye tracking has been used for a very long time in psychology, focusing on recording eye movements while reading. However, in the 1980’s researchers began to incorporate eye tracking into issues of human computer interaction. As technological tools such as the Internet, e-mail, and videoconferencing evolved into viable means of communication and information sharing during the 1990’s and beyond, researchers started using eye tracking in order to answer question about usability. Eye tracking technologies have been widely used as a proxy for users’ attention and the eye movement data gathered helps to understand where people focus attention, and in what order before they make a selection in their interactions with computer interfaces.

Goldberg and Kotval (1999) made a convincing argument for the use of eye-movement data in interface evaluation and usability testing. They claim that “Interface evaluation and usability testing are expensive, time-intensive exercises, often done with poorly documented standards and objectives. They are frequently qualitative, with poor reliability and sensitivity”. The motivating goal for their research work assessing eye movements as an indicator of interface usability was the provision of an improved tool for rapid and effective evaluation of graphical user interfaces.

They performed an analysis of eye movements (using interface designers and typical users) in order to assess the usability of an interface for a simple drawing tool. Comparing a ‘good’ interface with well-organized tool buttons to a ‘poor’ interface with a randomly organized set of tool buttons, the authors could show that the good interface resulted in shorter scan paths that cover smaller areas. The chief merit of this study was the introduction of a systematic classification of different measures for evaluating the usability of user interfaces based on eye movement data. These measures were grouped as follows: Measures of Processing (Number of fixations, Fixation Duration, Fixation / Saccade ratio, Scanpath Length, Scanpath Duration, Convex Hull Area, Spatial Density, Transition Matrix, Number of Saccades, Saccadic Amplitude), Other Measures (Backtrack, On-target / all-target fixations, Post-target fixations).

Usability Testing Software

Due to the escalating importance of HCI in software and Web design, some tools have been developed to assist the process of usability testing. Traditionally, usability testing involves the use of video cameras,
screen capture tools, microphones, and so forth. One of the main challenges lies in the integration of the media in various formats in a coherent way. Together with cheap computer hardware, usability testing software overcomes this problem by providing a coherent solution, in which Webcam, microphone and screen capture software operate concurrently under one system managed by the usability software. The output is seamless presentation of all media file which can be edited and annotated.

Typically, usability testing tools (hardware and software) consists of the following functionalities:

**User experience recording:**
The most common activity in usability testing is probably recording user performing tasks with the software or Web site under testing. This includes video recording (of the user), screen capture and voice recording (verbal protocol). With the usability testing tools, we can set the tasks needed to be performed, record the performance and administer questionnaires in an integrated environment.

**Observation and logging:**
Some usability tools support the connected of network computers thus allowing multiple observers to monitor the users and annotate the videos collaboratively. Furthermore, most tools can also log the mouse click and keyboard pressing.

**Analysis**

Although still limited to basic visualization, usability tools can support data analysis by calculating aggregate usability metrics for users or for tasks. For instance, we are able to quickly identify tasks with low usability metrics and thus focus on improving the design relevant to the tasks.

This reduces the amount of work and time significantly and it makes usability testing less costly.

**UTILIZATION AND APPLICATION OF HUMAN COMPUTER INTERACTION**

We now describe a number of key areas where HCI can have a practical application.

**Computer–Augmented Environments**

One application area in which HCI plays an important role is the computer-augmented environments, or commonly known as augmented reality or mixed reality. It refers to the combination of real world and computer-generated data visualization. In other words it is an environment which consists of both real world and virtual reality. For instance, a surgeon might be wearing goggles with computer generated medical data projected on it. The goggles are said to augment the information the surgeon can see in the real world through computer visualization. Therefore, it is not difficult to see the connection of augmented reality with ubiquitous computing and wearable computers.

Since its inception, augmented reality has had an impact on various application domains. The most common use is probably the support of complex tasks in which users need to perform a series of complicated actions while having access to large amount of information at the same time, such as surgery, assembly and navigation. Apart from these, augmented reality is also used for learning and training, such as flight and driving simulations.

Augmented reality implementation usually requires additional devices for input and output in order to integrate computer generated data into real world:
A Cave Automatic Virtual Environment multi-user, room-sized, high-resolution, 3D video and audio immersive environment in which the virtual reality environment is projected onto the walls. The user wearing a location sensor can move within the display boundaries, and the image will move with and surrounds the user.

A head-up display (HUD) is a transparent display that presents data without obstructing the user’s view. It is usually implemented on vehicles in which important information is projected directly in the driver’s viewing field. Thus the user does not need to shift attention between what is going on in the real world and the instrumental panel.

A head-mounted display is a display device, worn on the head or as part of a helmet that has a small display optic in front of one or both eyes.

Some of these devices have become commercially available and increasingly affordable. The challenge of HCI lies in the design of information visualisation which is not obtrusive to the users’ tasks.

**Computer-Based Learning**

A lot of effort has been put in coupling learning and technology to design effective and enjoyable learning. Various areas, namely e-learning, computer-based learning, serious games, etc have emerged, hoping to utilize the interactive power of computers to enhance teaching and learning experience. A myriad of design strategies have been proposed, implemented and evaluated, these include the early use of computer in presentation, drill and practice (the behaviourist paradigm), tutorials (cognitivist paradigm), games, story telling, simulations (constructivist paradigm), and so forth. As we progress from behaviorist to constructivist, we notice an explosion of user interface complexity. For instance, drill and practice programs usually consist on a couple of buttons (next, previous buttons, buttons for multiple choice, etc) while simulations could involve sophisticated visualization (outputs) and various user interface elements for manipulating parameters (input). Recently computer-based learning has moved from single user offline environments to online network spaces in which a massive number of users can interact with each other and form a virtual learner community. This social constructivist learning paradigm requires not only traditional usability treatment, but also sociability design in which the system includes not only the learning tools, but other sociability elements such as rules and division of labors.

**Information Visualization**

Information visualization is an area in HCI which can be related to many other areas such as augmented reality just described before. Most modern computer applications deal with visual outputs. Graphical user interface has almost entirely replaced command-based interaction in many domains. Information visualization can be defined as “the use of computer supported, interaction, visual representations of abstract data to amplify cognition” (Shneiderman, 1992). To amplify cognition means that visualization shifts cognitive loads to the perceptual system, thus expanding working memory and information storage.

Visualization provides a more perceptually intuitive way of viewing raw data, thus allowing users to identify relevant patterns which would not have been identified in raw data.

Therefore, it has a huge impact on many applications domains, ranging from engineering, education, various fields in science, and so forth.

In HCI, the most obvious application is the use of visualization is in the design of graphical user interface that allows more intuitive interaction between human and computers. Various innovative interaction styles have been developed such as WIMP (window, icon, menu, pointing device) which is so familiar in today’s software. Three-dimensional graphics are also emerging although currently they
are mostly used in computer games and computer-aided design. One recent example of 3D graphical interface is the new windows navigation and management known as Windows Flip 3D in Windows Vista which allows the user to easily identify and switch to another open window by displaying 3D snapshot thumbnail preview of all windows in stack.

Today, Information visualization is not only about creating graphical displays of complex information structures. It contributes to a broader range of social and collaborative activities. Recently, visualization techniques have been applied on social data to support social interaction, particularly in CMC. This area is known as social visualization by (Donath, Karahalios, & Viégas, 1999). Other technique such as social network analysis has also become increasingly important in visualization social data.

Other areas where HCI plays an important role include: Intelligent and agent systems; Interaction design; Interaction through wireless communication networks; Interfaces for distributed environments; Multimedia design; Non–verbal interfaces; Speech and natural language interfaces; Support for creativity; Tangible user interfaces; User interface development environments and User support systems.

ORGANISATIONAL AND SOCIAL IMPLICATIONS OF HUMAN COMPUTER INTERACTION

The application of HCI can have an effect of organisational and social dimensions. For example, the area of computer supported collaborative work (CSCW) explores the effect the introduction of technology can have an effect on the organisational structure and the way of work of companies and organisations. Similarly the study of how we use technology to communicate with each other is gaining strong interest in the HCI research community.

The expansion of the Internet has resulted in an increase in the usefulness of Computer Mediated Communication (CMC) and the popularity of online communities. It is estimated that 25% of Internet users have participated in chat rooms or online discussions (Madden & Rainie, 2003).

It is by now no secret how vital the Internet was, is, and will continue to be in our lives. One of the most important characteristics of this medium is the opportunities it offers for human-human communication through computer networks. As Metcalfe (1992) points out, communication is the Internet’s most important asset. E-mail is just one of the many modes of communication that can occur through the use of computers. Jones (1995) points out that through communication services, like the Internet, Usenet and bulletin boards, online communication has for many people supplanted the postal service, telephone and even the fax machine. All these applications where the computer is used to mediate communication are called Computer-Mediated Communication (CMC).

December (1997) defines CMC as “the process by which people create, exchange, and perceive information using networked telecommunications systems (or non-networked computers) that facilitate encoding, transmitting, and decoding messages”. He emphasizes that studies of CMC view this process from different interdisciplinary theoretical perspectives (social, cognitive/psychological, linguistic, cultural, technical, political) and often draw from fields such diverse as human communication, rhetoric and composition, media studies, human-computer interaction, journalism, telecommunications, computer science, technical communication and information studies.

Online communities emerge through the use of CMC applications. The term online community is multidisciplinary in nature, means different things to different people, and is slippery to define (Preece, 2000). For purposes of a general understanding of what online communities are, Rheingold’s definition of online communities is presented:
Online communities are also often referred to as cyber societies, cyber communities, Web groups, virtual communities, Web communities, virtual social networks and e-communities among several others.

The cyberspace is the new frontier in social relationships, and people are using the Internet to make friends, colleagues, lovers, as well as enemies (Suler, 2004). As Korzeny pointed out, even as early as 1978, online communities are formed around interests and not physical proximity (Korzeny, 1978). In general, what brings people together in an online community is common interests such as hobbies, ethnicity, education, beliefs. As Wallace (1999) points out, meeting in online communities eliminates prejudging based on someone’s appearance, and thus people with similar attitudes and ideas are attracted to each other.

It is estimated that as of September 2002 there are over 600 million people online (Nua Internet Surveys, 2004). The emergence of the so-called “global village” was predicted years ago (McLuhan, 1964) as a result of television and satellite technologies. However, it is argued by Fortner (1993) that “global metropolis” is a more representative term (Choi & Danowski, 2002). If one takes into account that the estimated world population of 2002 was 6.2 billion (U.S. Census Bureau, 2004), then the online population is nearly 10% of the world population – a significant percentage which must be taken into account when analyzing online communities. In most online communities, time, distance and availability are no longer disseminating factors. Given that the same individual may be part of several different and numerous online communities, it is obvious why online communities keep increasing in numbers, size and popularity.

Preece et al. (2002) states that an online community consists of people, a shared purpose, policies and computer systems. She identifies the following member roles:

- **Moderators and mediators**: who guide discussions/serve as arbiters
- **Professional commentators**: who give opinions/guide discussions
- **Provocateurs**: who provoke
- **General Participants**: who contribute to discussions
- **Lurkers**: who silently observe

CMC has its benefits as well as its limitations. For instance, CMC discussions are often potentially richer than face-to-face discussions. However, users with poor writing skills may be at a disadvantage when using text-based CMC (Scotcit, 2003).

**Examples of CMC and Online Communities**

Examples of CMC include asynchronous communication like e-mail and bulletin boards; synchronous communication like chatting; and information manipulation, retrieval and storage through computers and electronic databases (Ferris, 1997).

When it comes to Web site designers, choosing which CMC to employ (for instance, forum or chat-room) is not a matter of luck or randomness. Selecting the right CMC tool depends on a lot of factors. For example, in the case of e-learning, the choice of the appropriate mode of CMC will be made by asking and answering questions such as (Bates, 1995; CAP, 2004; Heeren, 1996; Resier and Gagne, 1983):
• Are the users spread across time zones? Can all participants meet at the same time?
• Do the users have access to the necessary equipment?
• What is the role of CMC in the course?
• Are the users good readers/writers?
• Is the activities time independent?
• How much control is allowed to the students?

Audio-conferencing is a real time communication mechanism, since the communication happens synchronously. Depending on the application, text chat and graphics may also be supported. Videoconferencing, like audio-conferencing, offers a useful mode of communication, but has the added benefit of being able to also see the participants, instead of just hearing them.

IRC and chats also support synchronous communication, since they enable the users to carry out conversations through the use of text messaging. MUDs build on chats by providing avatars and graphical environments where the users can engage in interactive fantasy games (Preece, 2000).

World Wide Web web sites are usually asynchronous, providing community information and links to other sites, but sometimes also have synchronous software, like chats, embedded in them (Preece, 2000).

E-mail is an asynchronous mode of communication usually in the form of text. However, the ability to add attachments to e-mail messages makes it possible for audio, video and graphics to be used also. Voice mail is an expansion of e-mail whereby users may record themselves speaking a message and then send that voice file to their contact, instead of typing it. Newsgroups, like e-mail, provide an asynchronous mode of communication, but unlike e-mail where the messages come to the users, it is a “pull” technology meaning the users must go to the UseNet groups themselves (Preece, 2000). Finally, discussion boards, also referred to as forums or bulletin boards provide an asynchronous mode of communication where the users post messages for others to see and respond at their own time.

Preece (2000) describes in detail some of these types of CMC and their different characteristics. In this chapter, we focus on Wiki and Online Virtual Game Environments which in our view provide a set of new, novel modes of communication and online community building systems.

Analyzing Online Communities: Frameworks and Methodologies

There are various aspects and attributes of CMC that can be studied to help us better understand online communities. For instance, the analysis of the frequency of exchanged messages and the formation of social networks, or, the analysis of the content of the exchanged messages and the formation of virtual communities. To achieve such an analysis a number of theoretical frameworks have been developed and proposed. For example Henri (1992) provides an analytical model for cognitive skills that can be used to analyze the process of learning within messages exchanged between students of various online e-learning communities. Mason’s work (1991) provides descriptive methodologies using both quantitative and qualitative analysis. Furthermore, five phases of interaction analysis are identified in Gunawardena et al.’s model (1997):

• Sharing/Comparing of Information
• The Discovery and Exploration of Dissonance or Inconsistency among Ideas, Concepts or Statements
• Negotiation of Meaning/Co-Construction of Knowledge
• Testing and Modification of Proposed Synthesis or Co-Construction
• Agreement Statement(s)/Applications of Newly Constructed Meaning
In this section we provide a description of some of the most commonly used online community evaluation techniques as well as their weaknesses and strengths.

**Studying CMC**

All of the already mentioned methods can be used to evaluate the usability and accessibility of the online community. But, apart from usability and accessibility we often also want to evaluate the user experience and the sociability of the interface. We describe below some of the methods that can assist us in this:

**Personas**

Findings from interviews and questionnaires can be further used as a basis for developing user profiles using personas. A persona is a precise description of the user of a system, and of what he/she wishes to accomplish. (Cooper, 1999). The specific purpose of a persona is to serve as a tool for software and product design. Although personas are not real people, they represent them throughout the design stage and are best based on real data collected through query based techniques.

Personas are rich in details, include name, social history and goals, and are synthesized from findings through the use of query based techniques with real people (Cooper, 1999). The technique takes user characteristics into account and creates a concrete profile of the typical user (Cooper, 1999).

For online communities, personas can be used to better understand the participants of the community and their background. Personas can also be used as a supplement to Social Network Analysis (described later in this chapter) to get a greater overview of the characteristics of key participants of a community. Using personas, Web developers gain a more complete picture of their prospective and/or current users and are able to design the interfaces and functionality of their systems, to be more personalized and suited for the communication of the members of their online communities.

Advantages of personas include: can be used to create user scenarios; can be anonymous protecting user privacy; represent the user stereotypes and characteristics.

Disadvantages of personas include: if not enough personas are used, users are forced to fall into a certain persona type which might now accurately represent them; time-consuming.

**Log Analysis**

A log, also referred to as Web-log, server log or log-file is in the form of a text file and is used to track the users’ interactions with the computer system they are using. The types of interaction recorded include key presses, device movements and other information about the user activities. The data is collected and analyzed using specialized software tools and the range of data collected depends on the log settings. Logs are also time stamped and can be used to calculate how long a user spends on a particular task or how long a user is lingered in a certain part of the Web site (Preece, Rogers & Sharp, 2002). In addition, an analysis of the server logs can help us find out: when people visited the site, the areas they navigated, the length of their visit, the frequency of their visits, their navigation patterns, from where they are connected and details about the computer they are using.

Log analysis is a useful and easy to use tool when analyzing online communities. For example, someone can use log analysis to answer more accurately questions like student attendance of an online learning community. Furthermore, logs can identify the Web pages users spend more time viewing, and also the paths that they used. This helps identify the navigation problems of the Web site, but also gives
a visualization of the users’ activities in the virtual communities. For instance, in the case of e-learning communities, the log files will show which students are active in the CMC postings even if they are not active participants (few postings themselves), but just observing the conversations. Preece (2003) notes that data logging does not interrupt the community, while at the same time can be used to examine mass interaction.

Advantages of Logs (Preece et al., 2002): helps evaluators analyze users behavior; helps evaluators understand how users worked on specific tasks; it is unobtrusive; large volumes of data can be logged automatically.

Disadvantages of Logs (Preece et al., 2002): powerful tools are needed to explore and analyze the data quantitatively and qualitatively; user privacy issues.

Content and Textual Analysis

Content analysis is an approach to understanding the processes that participants engage in as they exchange messages. There have been several frameworks created for studying the content of messages exchanged in online communities. Examples include work from Archer, Garrison, Anderson & Rourke (2001), Gunawardena, Lowe, and Anderson’s (1997) model for examining the social construction of knowledge in computer conferencing, Henri’s (1992) content analysis model and Fahy et al.’s (2001) Transcript Analysis Tool (TAT) which is described in more detail below.

The TAT focuses on the content and interaction patterns at the component level of the transcript (Fahy et al., 2001). After a lengthy experience with other transcript tools and reviews of previous studies Fahy et al. (2001), chose to adapt Zhu’s (1996) analytical model for the TAT. Zhu’s model (1996) examines the forms of electronic interaction and discourse, the forms of participation and the direction of participant interaction in computer conferences. The TAT also contains echoes of Vygotskian theory, primarily those dealing with collaborative sense making, social negotiation and proximal development (Cook & Ralston, 2003). The TAT developers have come up with the following strategic decisions (Fahy, 2003): The sentence is the unit of analysis; the TAT is the method of analysis; interaction is the criterion for judging conference success and topical progression (types and patterns).

The TAT was designed to permit transcript content to be coded reliably and efficiently (Fahy et al., 2001), while the advantages of TAT are (Fahy, 2003; Cook & Ralston, 2003; Fahy et al., 2001): It reveals interaction patterns that are useful in assessing different communication styles and online behavioral preferences among participants; It recognizes the complexity of e-conferences and measures the intensity of interaction; It enables the processes occurring within the conferences to be noted and recorded; It probes beyond superficial systems data, which mask the actual patterns of discussion; It relates usefully to other work in the area; It discriminates among the types of sentences within the transcript; It reflects the importance of both social and task-related content and outcomes in transcript analysis research.

The unit of analysis of the TAT is the sentence. In the case of highly elaborated sentences, the units of analysis can be independent clauses which, punctuated differently, could be sentences (Fahy, 2003). Fahy et al. (2001), have concluded that the selection of message-level units of analysis might partially explain problematic results that numerous researchers have had with previous transcript analysis work. They also believe that the finer granularity of sentence-level analysis results in several advantages (Fahy, 2003; Ridley & Avery, 1979): Reliability; Ability to detect and describe the nature of the widely varying social interaction, and differences in networking pattern, in the interactive behavior of an online community, including measures of social network density and intensity; Confirmation of gender associations in epistolary/expository interaction patterns, and in the use of linguistic qualifiers and intensifiers.
Social Network Analysis (SNA)

“Social Network Analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organisations, computers or other information/knowledge processing entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships” (Krebs, 2004, pp.1). Preece (2000) adds that it provides a philosophy and a set of techniques for understanding how people and groups relate to each other, and has been used extensively by sociologists (Wellman, 1982; Wellman 1992), communication researchers (Rice, 1994; Rice et al., 1990) and others. Analysts use SNA to determine if a network is tightly bounded, diversified or constricted; to find its density and clustering; and to study how the behavior of network members is affected by their positions and connections (Garton, Haythornthwaite & Wellman, 1997; Wellman, 1997; Henneman, 1998; Scott, 2000). Network researchers have developed a set of theoretical perspectives of network analysis. Some of these are (Bargotti, 2002):

- Focus on relationships between actors than the attributes of actors.
- Sense of interdependence: a molecular rather atomistic view.
- Structure affects substantive outcomes.
- Emergent effects.

“The aim of social network analysis is to describe why people communicate individually or in groups” (Preece, 2000, pp. 183), while the goals of SNA are (Dekker, 2002):

- To visualize relationships/communication between people and/or groups using diagrams.
- To study the factors which influence relationships and the correlations between them?
- To draw out implications of the relational data, including bottlenecks.
- To make recommendations to improve communication and workflow in an organisation.

Preece (2002) and Beidernikl & Paier (2003) list the following as the limitations of SNA:

- More theory that speaks directly to developers of online communities is needed
- The data collected may be personal or private.
- The analysis of the data is quantitative and specific to the particular network, while common survey data are qualitative and generalize answers on the parent population.

It is also worth pointing out that network analysis is concerned about dyadic attributes between pairs of actors (like kinship, roles, and actions), while social science is concerned with monadic attributes of the actor (like age, sex, and income).

There are two approaches to SNA:

**Ego-centered analysis**: Focuses on the individual as opposed to the whole network, and only a random sample of network population is normally involved (Zaphiris, Zacharia, & Rajasekaran, 2003). The data collected can be analyzed using standard computer packages for statistical analysis like SAS and SPSS (Garton, Haythornthwaite, & Wellman, 1997).
Whole network analysis: The whole population of the network is surveyed and this facilitates conceptualization of the complete network (Zaphiris et al., 2003). The data collected can be analyzed using microcomputer programs like UCINET and Krackplot (Garton et al., 1997).

The following are important units of analysis and concepts of SNA (Garton et al., 1997; Wellman, 1982; Hanneman, 2001; Zaphiris et al, 2003; Wellman, 1992):

- Nodes: The actors or subjects of study.
- Relations: The strands between actors. They are characterized by content, direction, and strength.
- Ties: Connect a pair of actors by one or more relations.
- Multiplexity: The more relations in a tie, the more multiplex the tie is.
- Composition: This is derived from the social attributes of both participants.
- Range: The size and heterogeneity of the social networks.
- Centrality: Measures who is central (powerful) or isolated in networks.
- Roles: Network roles are suggested by similarities in network members’ behavior.
- Density: The number of actual ties in a network compare to the total amount of ties that the network can theoretically support.
- Reachability: In order to be reachable, connections that can be traced from the source to the required actor must exit.
- Distance: The number of actors that information has to pass through to connect the one actor with another in the network.
- Cliques: Sub-sets of actors in a network, who are more closely tied to each other than to the other actor who are not part of the subset.

Social Network Analysis is a very valuable technique when it comes to analyzing online communities as it can provide a visual presentation of the community and more importantly it can provide us with qualitative and quantitative measures of the dynamics of the community.

EMERGING TRENDS IN HUMAN COMPUTER INTERACTION

Computer Game, Play and Immersion

Researchers from various areas have begun exploring the role of emotion in human activities. This research has proven to yield important results that not only deepen our understanding on human emotion, but potentially improves the design and development of artifacts. This goes without saying that the topic of emotion is becoming increasingly important in HCI. Probably one of the most studied domains in emotional design is the computer game.

Computer games have been widely recognized as the software which is potentially addictive due to their capability to engage and immerse the players for hours of game play. Therefore, it can be claimed that emotion is vital to maintain players’ focus and to create enjoyable experiences.

Unlike work-oriented activity (and thus work-oriented software that mediate them), game play traditionally does not focus on productive outcome. In this case, usability of game play cannot be understood in a traditional sense. Already new metrics of usability (or playability) have been proposed for designing playful artifacts/software.
Some effort has also been taken to incorporate game elements into productive activities. For instance, a project has been carried out to use games to label the contents of images meaningfully. Others have used games for education and training. An area of research, known as serious games, is expanding quickly to study productive games.

Recently, playability design has undergone a major transformation as games are becoming increasingly collaborative with the emergence of massively multiplayer online role-playing games (MMORPGs).

These games are becoming one of the most interesting interactive media of computer-mediated communication and networked activity environments (Taylor, 2002). Understanding the pattern of participation in these game communities is crucial, as these virtual communities function as a major mechanism of socialization of the players.

Some usability studies have shown that MMORPG design should incorporate what is known as sociability. For instance research has found that game locations can be designed to encourage different styles of social interactions

**CMC, Online Community and Social Computing**

Secondly, with the development of high performance computer networks, the use of computers is becoming increasingly collaborative. Therefore, topics such as CSCW, CMC, or social computing in general will continue to thrive.

One of the studied issues in this research topic is the use of non-verbal cues in the CMC technology. Some newer CMC media have attempted to address this issue through the implementation of video streaming, although a more common way is the use of “emoticons” that allows the users to express their emotions through graphical icons that show symbolic facial expressions.

Identity and anonymity is another popular topic regarding CMC since it is possible for users to hide their real identity simply by choosing a different name. In most cases, even the administrators do not know their real identity. Furthermore, users may treat CMC as an experimental environment and thus “play” with their identity and try out different personalities (Turkle, 1995).

Some studies have been carried out to compare the nature of friendship in conventional CMC and offline communication. Wellman and Gulia (1999) for instance found that people have a higher number of online friends because it is easier to make friends online than offline. However, the quality of the relationship is weaker in an online setting. Although Walther (1995) contended that it is possible to develop strong online friendship, it takes longer than in offline communication.

**3D CMC and Virtual World**

Conventional CMC technologies are mainly text-based (Pfeil, 2007). Although the fundamental technology of CMC has not changed in the last decade (Preece & Maloney-Krichmar, 2003), the way of using it for human-human communication has evolved considerably, from e-mail and synchronous chatting to online forum, blogs and Wikis.

New technologies such as 3D Virtual spaces, such as MMORPGs and Second Life (Linden Lab, 2003) offer a much richer form of social interactions where users can not only communicate through text-based chat, but also interact with each other through virtual artefacts. Furthermore, these technologies not only mediate social interaction, but also support group formation, which leads to community building. As such, user interaction is considerably more complicated than the conventional CMC tool.

These 3D worlds often feature a large number of virtual locations users can visit and a variety of virtual artefacts users can use to interact with, with each other. In some cases, they can even construct
new artefacts and locations utilising simple modelling tools and scripting languages. Whilst some 3D virtual worlds are designed with game-like goal structures that impose obstacles or challenges, some are completely open, meaning that the users are free to do as they please.

Although sociability issues of conventional CMC are well studied and documented, we have very little understanding on social interactions in 3D CMC. Therefore, it is worth investigating user activities in such environments in order to cast some light on the group formation process and other sociability issues.

It might be potentially more challenging in researching this 3D CMC, in which communication takes place both through texts and other “virtual actions” users can perform with their 3D avatars. Unlike the avatar in conventional CMC which is often a static graphical or animated representation of the user, in 3D CMC, the avatar can interact with other avatars directly in the virtual space. A 3D avatar can perform a wide range of actions on the 3D world and other avatars. For instance, it is not uncommon that avatars can hug, kiss or wave to each other. There is also research on the facial expression of 3D avatars (Clarkson et al., 2001), implemented in the 3D CMC context with the intention to enhance non-verbal communication. Moreover, given the fantasy theme of some 3D CMC environments, new sets of rules for virtual communication which are completely different from physical communication might arise. This is worth investigating as well as groups open operate within the boundary of norms and rules that emerge through user interaction.

**Ubiquitous Computing**

Another exciting future trend in HCI is the emergence of ubiquitous computing, in which information processing is thoroughly diffused into objects and experiences of everyday life. In another word, computers are disappearing. Users are no longer consciously engaged in using the computers. Instead they are operating these devices which are so well integrated into artifacts of everyday activities without being aware of using the computers.

Perhaps the most obvious example is the mobile phone, and indeed mobile computing has witnessed an explosion of research interest within and beyond HCI community.

Other less obvious examples could include computerized refrigerators which are able to detect their contents, plan and recommend a variety of recipes, automatically shop according to the users’ needs.

The focus of ubiquitous computing from the point of view of HCI suggests a shift from tool-focused design to activity-focused design. The primary objective is thus to design tools which can be seamlessly mediate everyday activities without interfering with users’ tasks. One such area is wearable computing which has the potential to support human cognitions, facilitate creativity and communication. Unlike mobile devices, wearable computers are attached to human, thus reducing the possibility of interruption or displacement.

**Novel Interaction Techniques**

With the integration of the computers into everyday life, there is a call for new interaction techniques as keyboards and mice are no longer sufficient as input devices.

Techniques such as natural language processing (NLP), a sub area of artificial intelligent, have been incorporated into the design of computer devices to recognize speeches, thus making interaction more intuitive. This automatic speech recognition can not only convert spoken words into machine readable strings, but also recognize and identify the person who speaks. Already, it has some impacts on our everyday life particularly on disable people who suffer from motor control impairment.
Perhaps a more ambitious application of NLP is to go beyond text-speech or speech-text conversion to actually understand human natural language. The research goal in this area is to develop a machine that can engage in meaningful conversation with human. Although still in its infancy, it is certainly a promising area which will benefit HCI greatly.

Another interesting interaction mode is what is known as haptic interaction which draws largely from the infamous interaction style in HCI, direct manipulation. The idea is to provide the users with greater freedom of manipulation that includes touch and feel. This technique aims to effectively bridge the gap between tangible interfaces and virtual/intangible representations. One example is the augmentation of everyday objects in which digital data is superimposed on physical objects which can be used to manipulate the digital data.

Recent development has also begun to utilize other parts of our body for interaction. Eye tracking for instance, which has been used extensively in research in psychology, cognitive linguistic and other areas, has been adopted in HCI research to study usability of Web sites and other software. Recently, its use in HCI has expanded to be applied as an interaction technique. Eye tracking system is now being used especially for the disabled to surf the Web, e-mail, and so forth with eye movements.

**Accessibility and Universal Design**

Finally we believe that Endeavour in designing assessable technology will continue to be a main focus in HCI. One major accessibility research is the Web design since the Web has rapidly become more and more pervasive in almost everybody’s lives.

The challenge lies in the creation of resources that can be used by the widest spectrum of potential visitors rather than an idealized “average,” there is a need to consider universality of Web delivery.

The idea accessibility has gone beyond the basic concept of “designing for disability” to universal design, allowing access by everyone.

This requires consideration of the needs and requirements of individuals with disabilities, older persons, people for whom English is a second language, people whose cultures and backgrounds are dissimilar to those of Web developers, and those using outdated hardware and software, to name a few.

For that reason, the key objective of this research direction is to look at various aspects of universal Web design and evaluation in a new perspective by focusing on the user aspect of universal Web design and interaction and to present the wide range of advanced technology that can help disadvantaged users get access to Web information.

The same idea also applies to software design. For instance, child-computer interaction has emerged as a sub-field of HCI that deals with designing particularly multimedia learning software for children. Other active areas of research include web design for elderly people, gender differences, cultural issues, and so forth.

All these efforts in accessibility design no longer contribute to the development of ease-of-use applications for specific user groups, but also improve interaction experiences for mainstream users in general.

**REFERENCES**


Fahy, P. J. (2003). Indicators of support in online interaction. *International Review of Research in Open and Distance Learning, 4*(1).


of Human-Computer Studies, 55, 4, 485-502.


About the Editor

Panayiotis Zaphiris is a reader in human-computer interaction at the Centre for HCI Design. He got his B.Sc and M.Sc. from the University of Maryland College Park (USA). Panayiotis got his PhD (in human-computer interaction) from Wayne State University where he was also a research assistant at the Institute of Gerontology. Panayiotis’ research interests lie in HCI with an emphasis on inclusive design and social aspects of computing. He is also interested in Internet related research (Web usability, mathematical modelling of browsing behaviour in hierarchical online information systems, online communities, e-learning, Web based digital libraries and finally social network analysis of online human-to-human interactions). He has authored over 110 papers and four books in this area including a recently released book titled User-Centered Design of Online Learning Communities (Lambropoulos and Zaphiris, 2006). He is principal investigator for the AW-Model EPSRC/BBSRC project funded under the SPARC scheme. Panayiotis was the principal investigator on the JISC-funded “Usability Studies for JISC Services and Information Environment” and “Information Visualisation Foundation Study” projects which looked at usability and information visualisation in the context of online learning environments. Finally, he was the co-principal investigator on the largest ever conducted Web accessibility study (the DRC funded “An in-depth study of the current state of Web accessibility” project) that assessed at the accessibility of 1000 UK Web sites. More information about Panayiotis’ research can be found at: http://www.zaphiris.org

Chee Siang Ang is a research fellow in the Centre for Human-Computer Interaction (HCI) Design. He is interested in human interactions and social tendencies in the virtual world, particularly Second Life, from a sociological, psychological and HCI perspective. His main research interests include the psychology and sociology of computer games, virtual worlds or 3D Computer-Mediated Communication (CMC), learning theories particularly in gaming, digital media such as interactive narrative and simulation.
This section serves as the foundation for this exhaustive reference tool by addressing crucial theories essential to the understanding of human-computer interaction. Chapters found within these pages provide an excellent framework in which to position human-computer interaction within the field of information science and technology. Individual contributions provide overviews of ubiquitous computing, cognitive informatics, and sociotechnical theory, while also exploring critical stumbling blocks of this field. Within this introductory section, the reader can learn and choose from a compendium of expert research on the elemental theories underscoring the research and application of human-computer interaction.
Chapter 1.1
Introduction to Ubiquitous Computing

Max Mühlhäuser
Technische Universität Darmstadt, Germany

Iryna Gurevych
Technische Universität Darmstadt, Germany

A BRIEF HISTORY OF UBIQUITOUS COMPUTING

Mark Weiser

The term ubiquitous computing was coined and introduced by the late Mark Weiser (1952-1999). He worked at the Xerox Palo Alto Research Center (PARC, now an independent organization). PARC was more or less the birthplace of many developments that marked the PC era, such as the mouse, windows-based user interfaces, and the desktop metaphor (note that Xerox STAR preceded the Apple Lisa, which again preceded Microsoft Windows), laser printers, many concepts of computer supported cooperative work (CSCW) and media spaces, and much more. This success is contributed (among other reasons) to the fact that PARC managed to integrate technology research and humanities research (computer science and “human factors” in particular) in a truly interdisciplinary way. This is important to bear in mind since a considerable number of publications argue that the difference between UC and Ambient Intelligence was the more technology/networks-centered focus of the former and the more interdisciplinary nature of the latter that considered human and societal factors. We do not agree with this argument, in particular due to the nature of the original UC research at PARC—and the fact that quite a number of UC research labs worldwide try to follow the PARC mindset. Indeed, Mark Weiser concentrated so much on user aspects that quite a number of his first prototypes were mere mockups: during corresponding user studies, users had to imagine the technology side of the devices investigated and focus on use cases, ideal form factors and desired features, integration into a pretend intelligent environment, and so forth.
Weiser’s Vision of UC


Maybe the most frequently cited quotation from this article reads as follows: “The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.” This was Mark’s vision for the final step in a development away from “standard PCs”, towards a proliferation and diversification of interconnected computer-based devices. A deeper understanding of Mark Weiser’s visions can be drawn from his position towards three dominant, maybe overhyped trends in computer science at his time: virtual reality, artificial intelligence, and user agents. With a good sense for how to raise public attention, Mark criticized these three trends as leading in the wrong direction and positioned UC as a kind of “opposite trend”. We will follow Mark’s arguments for a short while and take a less dramatic view afterwards.

UC vs. Virtual Reality (VR)

According to Mark, VR “brings the world into the computer”, whereas UC “brings the computer into the world”. What he meant was that VR technology is generally based on elaborate models of an existing or imagined (excerpt of the) world. This model contains not only 3D (geometric) aspects but many more static and dynamic descriptions of what is modeled. For instance, digital mockups of cars have been pushed to the point of simulating crash tests based on the car /obstacle geometry, static, and dynamic material characteristics, laws of physics, and so forth. As the sophistication of models grows, more and more aspects of the world are entered into the computer, finally almost everything happens in the virtual space and even the human becomes a peripheral device for the computer, attached via data gloves and head-mounted displays. Mark Weiser criticized mainly the central and peripheral roles of computers and humans, respectively. He proposed to follow the UC vision in order to *invert* these roles: by abandoning the central role of computers and by embedding them in the environment (in physical objects, in particular), room is made for the human in the center. In this context, he used the term “embodied virtuality” as a synonym for UC. The cartoons in Figure 1 were made by Mark Weiser and provided by courtesy of PARC, the Palo Alto Research Center, Inc.

UC vs. Artificial Intelligence (AI)

In essence, Mark Weiser criticized the overly high expectations associated with AI in the 1980’s. In the late 1980’s and early 1990’s, that is, at the time when he developed his UC vision, AI research had to undergo a serious confidence crisis. The term AI had not been associated with a commonly accepted, reasonably realistic definition, so that the association with human intelligence (or the human brain) was destined to lead to disappointments. The AI hype had provided researchers with considerable funds—but only for a while. Mark Weiser proposed to take a different approach towards a higher level of sophistication of computer-based solutions (which had been the goal of AI at large). He considered it a more reasonable objective to concentrate on small subsets of “intelligent behavior” and to dedicate each computer to such a subset. Higher sophistication would be fostered by interconnecting the special-purpose computers and by making them cooperate. This reasoning lead to the term *smart*, considered more modest than the term *intelligent*. Sensor technology plays an important role in dedicating computers to a small subset of “understanding the world around us” (a key element of intelligent behavior). By widely deploying and interconnect-
ing sensor-based tiny computers, one would be able to integrate environmental data (location, temperature, lighting, movement, etc.) and use this information to produce smart behavior of computers and computerized physical objects.

**UC vs. User Agents (UA)**

In contrast to virtual reality and artificial intelligence, the term *user agent* is not very prominent in the general public. At the time referred to, UAs were thought as intelligent intermediaries between the user and the computer world, that is, as an approach towards increased ease-of-use or better human-computer interaction. User agents were often compared to the common perception of British butlers who are very discreet and unobtrusive, but always at disposal and extremely knowledgeable about the wishes and habits of their employers. Following this analogy, UAs were installed as autonomous software components between applications and users, inspecting and learning from the user-software application. Mark Weiser challenged five requirements usually derived from this analogy for user agents and proposed UA as a better alternative for the first three; as to the last two, he judged the necessary base technology as immature:

1. UAs were supposed to give advice to their users based on what they had learned. Mark Weiser asked, in essence, why they would not do the job themselves—a promise that UC should fulfill;
2. UAs were supposed to obey the user, for example, by applying planning algorithms to basic operations with the aim to fulfill the goals set by a user. In contrast to this approach, UC was intended to behave rather proactively, that is, to propose and even act in advance as opposed to reacting on command;
3. A third widespread requirement suggested that UAs should intercept the user-application interface. UC in contrast should be more radical and take over the interaction or carry out functions on its own—an approach presumed by Mark Weiser to be the only viable one if humans were to be surrounded by hundreds of computers;
4. A basic assumption about UAs was that they would listen to the (interactions of) the user. Mark Weiser considered natural language processing technology and speech recognition technology at his time to be far too immature to promise satisfying results in this respect;
5. UAs should learn the users’ preferences, wishes, and so forth by observation. Again, the necessary (machine learning) technology was judged to be too immature to live up to this promise.

![Figure 1. Mark Weiser’s cartoons about UC vs. virtual reality](image-url)
We will resume the VR / AI / UA discussion in the next large section.

Mark Weiser’s Three Key Devices

We want to finish this lengthy but still extremely compressed, much simplifying and abstracting treatment of Mark Weiser’s contributions by looking at three devices. These complementary UC devices were prototyped at his lab; investigated in the context of PARC’s typical creative, team-oriented setting, all three were thought as electronic replacements for the common “analog” information appliances.

The Xerox “Pad” can be considered to be the prototype and father of present PDA’s, introduced even before Apple’s Newton appeared in 1993. The initial concept was that of an electronic equivalent to “inch-size” information bearers, namely “PostIt Notes”: easy to create and to stick almost everywhere, available in large quantities. As the PDA analogy suggests, the prototypes had a lot more functionality than PostIt Notes—but were also a lot more expensive and cumbersome to handle by design (not only due to short and mid-term technology limitations).

The Xerox “Tab” can be considered to be the prototype and father of present Tablet PC’s. The analogy from the traditional world was that of a “foot-size” information bearer, namely a notebook or notepad. One may infer from the rather stalling market penetration of Tablet PC’s that technology is still not ready for mass market “Tabs” today, but one may also expect to find a pen centric, foot size, handheld computer to become very successful any time soon. An interesting facet of the original Tab concept was the idea that Tabs would in the future lay around for free use pretty much as one finds paper notebooks today, for example, as part of the complementary stationery offered to meeting participants.

The Xerox “Liveboard” was the prototype of present electronic whiteboards. A PARC spinoff company designed and marketed such boards, and today many companies like Calgary-based SmartTechnologies Inc. still sell such devices. Liveboards represented the “yard-size” information bearers in the family of cooperating devices for cooperating people. In contrast to many devices sold today, Liveboards supported multi-user input pretty early on.

The developments and studies conducted at Mark Weiser’s lab emphasized the combination of the three device types for computer supported cooperation, and cooperative knowledge work in particular.

While Mark Weiser was a truly outstanding visionary person with respect to predicting the future of hardware, that is, UC nodes (proliferation of worn and embedded networked devices, specialized instead of personal general-purpose computers, numbers by far exceeding the number of human users), two other people were more instrumental in generating awareness for the two remaining big challenges mentioned in the preface of this book, namely integrative cooperation and humane computing; the former of these challenges was emphasized by Kevin Kelly, the latter by Don Norman. A deeper analysis reveals that for the second aspect, humane computing, it is very difficult to argue about the true protagonists. Readers remember that Mark Weiser was actually placing a lot of emphasis on usability, by virtue of his education and mindset and in the context of the human focus of PARC. He also coined the exaggerated term “invisible” for mature technology. On the other hand, Don Norman was not advocating the humane computing challenge in all its facets yet. Nevertheless, we want to highlight him next as maybe the single most important advocate of this challenge.

The Book Out of Control by Kevin Kelly

In 1994, K. Kelly published a book entitled Out of Control. The thoughts expressed by Kelly were an excellent complement to Mark Weiser’s
publications. While the latter emphasized the emergence of networked small “neuron like” (i.e., smart) UC nodes, Kelly emphasized the integrated whole that these neurons should form. His starting argument was the substantiated observation that the complexity of the made, that is, of human-made systems or technology, approached the complexity of the born, that is, of “nature-made” systems, such as human or biological organisms, human or biological societies (cf. ant colonies), and so forth.

This observation led to the obvious requirement to investigate the intrinsic principles and mechanisms of how the born organized, evolved, and so forth. By properly adopting these principles to “the made”, this complexity might be coped with. Research about the organization and evolution of the born should be particularly concerned with questions such as: how do they cope with errors, with change, with control, with goals, and so forth. For instance, beehives were found not to follow a controlling head (the queen bee does not fulfill this function), and it is often very difficult to discern primary from subordinate goals and to find out how goals of the whole are realized as goals of the individuals in a totally decentralized setting.

Kevin Kelly summarizes central findings and laws of nature several times with different foci. Therefore, it is not possible to list and discuss these partly conflicting findings here in detail. An incomplete list of perceived central laws “of God” reads as follows: (1) give away control: make individuals autonomous, endow them with responsible behavior as parts of the whole, (2) accept errors, even “build it in” as an essential means for selection and constant adaptation and optimization, (3) distribute control truly; that is, try to live with no central instance at all, (4) promote chunks of different kinds (e.g., hierarchies) for taming complexity, and (5) accept heterogeneity and disequilibrium as sound bases for survival.

The Book The Invisible Computer by Donald Norman

Don Norman emphasized the “humane computing” grand challenge described in the preface of this book. World renowned as an expert on usability and user-centered design, he published The Invisible Computer in 1999. He considered the usability problems of PC’s to be intrinsically related to their general-purpose nature and thus perceived the dawning UC era more as a chance than a risk for humane computing. The intrinsic usability problems that he attributed to PCs were rooted in two main anomalies, according to Don Norman: (1) PCs try to be all-purpose and all-user devices—a fact that makes them overly complex, and (2) PC’s are isolated and separated from daily work and life; truly intuitive use—in the context of known daily tasks—is therefore hardly possible. From this analysis, Norman derived various design guidelines, patterns, and methodological implications, which we will summarize again at an extremely coarse level:

1. He advocated UC nodes using the term “information appliances”: dedicated to a specific task or problem, they can be far simpler and more optimized;
2. He further advocated user-centered development: especially with a specific user group in mind, “information appliances” as described previously can be further tailored to optimally support their users;
3. Norman stated three key axioms, that is, basic goals to be pursued during design and development: simplicity (a drastic contrast to the epidemic “featurism” of PC software), versatility, and pleasurability as an often forgotten yet success critical factor;
4. As a cross-reference to the second big UC challenge (integrative cooperation), he advocated “families of appliances” that can be easily and very flexibly composed into systems.
History Revised

The preceding paragraphs are important to know for a deeper understanding of the mindset and roots of UC. However, about 15 years after the time when the corresponding arguments were exchanged, it is important to review them critically in the light of what has happened since. We will first revise the three “religious disputes” that Mark Weiser conducted against AI, VR, and UAs. To put the bottom line first, the word “versus” should rather be replaced by “and” today, meaning that the scientific disciplines mentioned should be (and have, mostly) reconciled:

As to UC and VR, specialized nodes in a global UC network can only contribute to a meaningful holistic purpose if models exist that help to cooperatively process the many specialist purposes of the UC nodes. In other words, we need the computer embedded into the world and the world embedded in the computer. Real Time Enterprises are a good example for very complex models—in this case, of enterprises—for which the large-scale deployment of UC technology provides online connectivity to the computers embedded into the world, that is, specialized nodes (appliances, smart labels, etc.). In this case, the complex models are usually not considered VR models, but they play the same role as VR models in Mark Weiser’s arguments. The progress made in the area of augmented reality is another excellent example of the benefit of reconciliation between UC and VR: in corresponding applications, real-world vision and virtual (graphical) worlds are tightly synchronized and overlaid.

As to UC and AI, Mark Weiser had not addressed the issue of how interconnected, smart, that is, “modest”, specialized nodes would be integrated into a sophisticated holistic solution. If the difference between AI and the functionality of a single smart UC node (e.g., temperature sensor) was comparable to the difference between a brain and a few neurons, then how can the equivalent of the transition (evolution) from five pounds of neurons to a well-functioning brain be achieved? Mark Weiser did not have a good answer to that question—such an answer would have “sounded like AI” anyway.

Today, there is still not a simple answer yet. The most sophisticated computer science technology is needed in order to meet the integration challenge of how to make a meaningful whole out of the interconnected UC nodes. However, the state of the art has advanced a lot and our understanding for what can be achieved and what not (in short term) has improved. For instance, socionic and bionic approaches have become recognized research areas. A mature set of methods and algorithms is taught in typical “Introduction to AI” classes today and has replaced the ill-defined, fuzzy former understanding of the area. Thus the boundaries between AI and computer science are more blurred than ever and their discussion is left to the public and press.

As to UC and UAs, remember that Mark Weiser considered UAs as “too little” in terms of what they attempted (at least too little for the UC world envisioned by him), yet “too much” in terms of what the underlying technology was able to provide. This left doubts about how the even more ambitious goals of UC could be met, namely active (proactive, autonomous, even responsible) rather than reactive (obeying) behavior. In other words, Mark Weiser was right when he advocated active as opposed to reactive behavior, but he had little to offer for getting there. Luckily, the technologies that he had then considered immature (e.g., speech processing, NLP, machine learning) have advanced a lot since.

All in all, Mark Weiser’s arguments from 15 years ago (1) provide a deep understanding of the field, (2) should be modified towards a more conciliatory attitude (in particular with respect to AI and VR / complex “world models”), and (3) have become more substantiated in certain respects since technology advancements make some of his more audacious assumptions more realistic (but most visions of his “opponents”, too). In other
words, Mark Weiser’s visions were and still are marking the research and developments made by the UC community. His concepts and predictions were accurate to a degree that was hardly paralleled by any other visionary person. Restrictions apply as to his overly drastic opposition to VR, AI, and UAs: some of the exaggerated promises of these were repeated by him in the UC context - right when he denounced the over-expectations raised by AI and UAs! VR and AI in particular should be reconciled with UC. Maybe Weiser underestimated the two grand challenges of the UC era, namely “integrative cooperation” and “humane computing”.

Kevin Kelly and Donald Norman emphasized these two challenges, respectively. Looking at the advancements in totally decentralized systems, Kelly’s promises can be evaluated as too extreme today: bionics social science inspired, and autonomous (or autonomous) computing have advanced a lot. However, two restrictions still apply: (1) less decentralized systems still prove to be extremely viable in daily operation—it will be hard for fully decentralized systems to really prove their superiority in practice; (2) system-wide goals must still be planned by some centralized authority and—to a certain extent manually—translated into methods for fully decentralized goal pursuit; evolution-like approaches that would generate optimization rules and their pursuit automatically in a fully decentralized systems are still hardly viable. As a consequence, the present book will not only describe the above-mentioned computing approaches in part “Scalability”, but also other aspects of scalability.

As to Don Norman, he was right to advocate simplicity as a primary and key challenge. However, he maybe underestimated the ‘humane computing’ problems associated with the nomadic characteristics and ‘integrative cooperation’ challenge of the UC era. The usability of the integrated whole that we advocate to build out of UC nodes is by far not automatically endowed with easy-to-use user interaction just because the participating appliances exhibit a high degree of usability. On the other hand, only the integration that is, federation of miniature appliances with large interaction devices (wall displays, room surround sound, etc.) may be able to provide the usability desired for an individual device.

As we conclude this section, we should not forget to mention that the UC era was of course not only marked by just three visionary people.

**TERMS AND SELECTED STANDARDS**

While there is a lot of agreement among researchers and practitioners worldwide that the third era of computing is dawning as the era of networked, worn/portable and embedded computers, there is not so much agreement about what to call that era. This fact is something of an obstacle, for instance for wider recognition in politics (the crowd does not scream the same name as one may put it). This situation is aggravated by the fact that partial issues and aspects of Ubiquitous Computing are also suffering from buzzword inflation. With this background in mind, one may understand why we list a considerable number of these buzzwords below and provide a short explanation, rather than swapping this issue out into a glossary alone. Knowledge of the following terms is indeed necessary for attaining a decent level of “UC literacy”.

**Synonyms for Ubiquitous Computing**

First, we want to look at the terms that describe—more or less—the third era of computing as introduced:

- **Post-PC era**: The root of this term is obvious, it describes ‘the era that comes after the second, that is, the PC era. We suggest avoiding this term since it points at what it
Pervasive computing: A distinction between the word ubiquitous and pervasive is difficult if not artificial. One could argue that the term pervasive eludes more to the process of penetration (i.e., to the verb pervade) whereas ubiquitous eludes more to the final state of this process. We suggest that pervasive computing and ubiquitous computing are synonyms, one (pervasive) being slightly more common in industry (its origin has been attributed to IBM), the other one (UC) being slightly more common in academia.

Ubiquitous computing: The term may be interpreted as “computers everywhere”. We are using it as the notion for the third era of computing throughout the book and prefer it, among others, because we try to fight buzzword mania and dislike the invention of additional terms for a named concept. We therefore propose to stick to the first (reasonable) term invented and somewhat broadly accepted; since Mark Weiser is the first visionary person who sketched essential characteristics of the dawning era and since he invented the term UC, the question of what is the oldest well-known term should not be questionable.

Ambient intelligence: This term was invented in particular in the context of the European Union’s research framework programs (5, 6, 7). As a positive argument, one may say that the two words reflect the grand challenges of UC as stated in this book: ambient may be associated with the challenge of humane computing, making UC systems an integral part of our daily life. Intelligence may be interpreted as the challenge of integrative cooperation of the whole that consists of myriads of interconnected UC nodes. On the downside, one should remember that Mark Weiser had intentionally avoided the term “intelligence” due to the over-expectations that AI had raised. We suggest avoiding this term, too, because it is still burdened with these over-expectations and because it is still ill defined.

Disappearing / invisible / calm computing: All three terms are less common than UC and pervasive computing. Their roots have been discussed in the historical context above. Obviously, disappearing describes again a process while “invisible” describes a final state. “Calm” emphasizes hearing as opposed to vision like the other two. In any case, the terms “invisible” and “disappearing” are not very well chosen (despite our tribute to Don Norman) since computers and interfaces that have totally disappeared cannot be commanded or controlled by humans any more. Since we doubt that 100% satisfactory service to the user can be paid at all without leaving the customer, that is the user, the option to explicitly influence the service behavior, we consider the term misleading. We favor again Mark Weiser’s notion of computers that are so well interwoven with the fabric of our lives that we hardly notice them.

Mixed-mode systems: This is a term used to describe the heterogeneity of UC nodes, in contrast to the rather resource rich, general purpose PC’s of the last era. This term is even less common, but pops up every now and then like those previously discussed, and should not be used to describe UC as a whole since it emphasizes a particular aspect.

Tangible bits: This term has found some currency in the Netherlands and Japan, but remained rather uncommon in general. It refers mainly to the fact that networked computers are becoming part of the physical world.

Real time enterprise: This term has been explained in the preface of the book and is not
thought as a synonym for UC, but rather as a very important and cutting-edge application domain that may drive down the learning curve, that is, prices of UC hardware and solutions.

It was mentioned in the preface that some authors argued in favor of one or the other of the UC synonyms, saying that their choice was more far-reaching in time (the other ones being intermediate steps) or space (the other ones only comprising a subset of the relevant issues). However, we cannot follow these arguments, mainly because research labs and projects around the world work on the same subjects, some more advanced or holistic, some less ambitious or more specialized, carrying the names UC, pervasive computing, and ambient intelligence rather randomly.

Towards a Taxonomy of UC Nodes

Throughout this book, UC nodes will be categorized according to different aspects. In the context of reference architectures further below, we will emphasize the role of UC nodes in a holistic picture. In the present paragraph, we want to try categorizing them as devices. It should be noted that in the preface of the book, we already provided a preliminary, light weight introduction. The difference between carried (worn, portable) and encountered nodes was emphasized and four preliminary categories (wearables, sensors, appliances, and smart labels) were briefly described. It soon became clear that smart labels attached to goods must be distinguished again from those attached to humans, although the base technology may be the same.

In a second, more serious attempt to categorize UC nodes as device categories, we propose the following distinction (see Figure 2):

1. Devices attached to humans
   a. Devices carried: Here we further distinguish three subcategories: (1) mobile devices, synonymous with portable devices, contain rather general purpose computers and range from laptops via PDA’s to mobile phones and the like, (2) smart badges, that is, smart labels serve for identification, authentication and authorization of humans and possibly further purposes, and (3) body sensors of all kinds play an increasingly important role in particular in the fitness and health context;
   b. Devices worn: These wearables range from truly sophisticated, computer-augmented cloths and accessories to prototypes that are built from standard components (PDA in a holster with headset, etc.). A further categorization is not attempted since the spectrum is rather blurred;
   c. Devices implanted: while there is a lot of hype about implanted RFID tags and networked health implants, the many issues (e.g., health, privacy, or dependability) around the necessary device-environment communication have not permitted this category to become widespread.

2. Devices encountered
   a. Smart items denote computer-augmented physical objects. The terms “smart object” and “smart product” are used with subtle differences depending on the context to denote more sophisticated variants of smart items, such as smart items that proactively communicate with the users. We suggest treating smart items as the most general term and to distinguish the following subcategories: (1) smart tags as the least sophisticated variant: they can be considered to be mimicry for embedded computers: by attaching a smart tag to a physical object, a physically remote computer (often
in proximity, though) can take over some of the functionality that would be embedded otherwise. This approach opens the door for turning even the cheapest products into UC nodes. The term “smart label” is sometimes used synonymously; sometimes it is used as the comprehensive term for smart tags and smart badges (attached to humans, see earlier discussion). We suggest sticking to the term smart tag for the smart item sub-category described here; (2) networked sensor nodes, and (3) networked appliances denote the other subcategories of smart items. They were already introduced in the preface of this book.

b. **Smart environments** denote the surroundings of smart items, that is, the additional communication and compute power installed in order to turn an assembly of smart items into a local, meaningful whole.

The reader must be aware that all terms arranged in the taxonomy are not settled yet for a common understanding. For instance, one might argue whether a sensor network that computes context information for networked appliances and users should be considered a set of smart items (as we defined it) or a part of the smart environment. Nevertheless, we find it useful to associate a well-defined meaning with these terms and to apply it throughout the book (see Figure 2).

In addition, it should be noted that **smart environments** (with integrated smart items) constitute a particularly important research area—maybe because they permit researchers and project leaders to implement self-contained “little UC worlds” without a need for multiparty agreements about interoperability standards. In particular, “smart homes” were among the first subjects of investigation in the young history of UC. Prestigious projects in the smart home area were and are conducted by industry (Microsoft eHome, Philips AmbientIntelligence initiative, etc.) and academia (GeorgiaTech AwareHome, MIT House, etc.). HP made an early attempt to overcome the isolation of such incompatible islands by emphasizing standard middleware in the Cooltown project). Quite a number of projects about smart homes terminated without exciting results, not to the least due to insufficient business impact (note our argument in favor of Real Time Enterprises as a more promising subject). More recently, smart homes projects have focused on issues considered to be particularly promising, as was discussed in the preface to this book. Important areas comprise home security, energy conservation, home entertainment, and particu-
larly assisted living for the aging society—a topic considered particularly interesting in Europe (1 year prolongation of independent living saving about half a billion Euros in Germany alone). Renowned large-scale projects were carried out, for example, in Zwijndrecht (Belgium) and Tønsberg (Norway) in this respect.

A Few More Relevant Terms

A few more UC terms—and sometimes, corresponding concepts—are worth mentioning.

- **Smart dust** is a term used for sensor networks if the emphasis is on miniaturization and the concept is based on one-time deployment and zero maintenance. Environment data sensors are often cited as an example, the vision then is to deploy them, for instance, from an aircraft, and let them monitor the environment until they fail. Environment-friendly degradation is a major issue here, of course.

- **Things that think** was the name of an early UC project led by Nicholas Negroponte at the MIT media lab. Other authors have since hijacked the term.

- **Smart paper** denotes the vision of a display device that would exhibit characteristics comparable to traditional paper in terms of weight, robustness, readability, and so forth, and loadable with the content of newspapers, journals, books and so forth. It would help to save paper and revolutionize the press distribution channels and more. Many projects that were not even close to this vision had, and continue to have, the name “smart paper”.

- **Smart wallpaper** is a similar term to smart paper in that it extrapolates the above mentioned characteristics to wall-size devices.

- **Smart <you-name-it>:** virtually every noun has been associated with the attribute smart recently, not always alluding to the characteristics of UC nodes. For instance, smart materials are supposed to adapt to the context of use, with no IT involved. Most of the time though, smart <something> alludes to a physical object that has been augmented with an embedded computer.

- **The Internet of things** is a term favored by the press. It is not considered appropriate as a term for UC as a whole by the authors since it emphasizes the hardware side of UC as opposed to the human side, which was already described as crucial and as a major challenge (cf. humane computing). Most publications that favor this term concentrate on the two standards discussed in the following section.

The EPCglobal Standard

As mentioned at the beginning, we will only sketch two important standards in the UC context. Other standards are too unimportant, too immature, too specific (they might be treated in one of the focused parts of this book), or part of the background knowledge about well-established technology that this book cannot cover. The first standard to mention is EPCglobal and was mentioned in the preface of this book. As mentioned, it is meant to succeed the barcodes that encode the European article number or universal product code on current consumer products. The 96-bit Electronic Product Code EPC is usually stored on RFIDs (a subcategory of smart tags, as we can now say) and can be read:

- From a greater distance (e.g., 10m)
- With better reading accuracy
- With much less effort (e.g., en-passant by a RFID reader gate as opposed to carefully with line-of-sight connection by a barcode scanner)
- In bulk (RFID readers can read, for example, a hundred tags at once)
Since the EPC contains a 36-bit serial number, individual items can be tracked and traced. For instance, theft can be much more easily attributed to criminals, product life cycles can be recorded more accurately, product lots with manufacturing errors can be called back more specifically, etc. On the other hand, the serial number may in principle be used to trace an individual, too, if she carries around an RFID tagged product. This privacy issue has raised many concerns in recent years and amplified the decision of the whole sales and retail industry to focus on tagging their containers, palettes, cases, etc., for a start. So-called item level tagging is only envisioned for highly valuable goods initially; it may enter the mass market when tag prices and system costs have come down and after settling the privacy issues.

Figure 3 depicts the functioning of EPC smart tags in an overall IT infrastructure. In step 1, an EPC code is read from a product. In the example, each carton on the palette could contain a number of tagged products. The residual example would then explain the action for just one such tag. Usually prior to reading the tag, the system has already searched and discovered servers capable of ‘resolving’ certain ranges of EPC code. Based on the results of this discovery process, the appropriate ‘resolution node’, called an ONS server, is asked to resolve the EPC code, that is to translate it into a global Internet address where the relevant product information is actually stored. The product information is encoded in a standardized way, using the so-called product markup language PML, an XML derivate.

The second generation of RFID tags introduced in 2006 features improved bulk reading (hundreds of tags simultaneously), size and cost improvements. “Printable” tags have become common: these paper labels with embedded RFID chips can be custom imprinted with custom human-readable information. The chips themselves are not altered in the printer and they come with pre-assigned EPC codes from the manufacturer.

The OSGi Standard

The Open Services Gateway Initiative (OSGi) is an industry driven nonprofit consortium. OSGi standardized a Java virtual machine (JVM). This JVM can be considered a standardized virtual ‘computer’ that runs on any real computer and is capable of executing programs that are transmitted to it, so-called bundles. OSGi standardizes not only the format for bundles, but also the necessary protocols and procedures for authenticating and authorizing senders of bundles, for replacing and updating bundles (remote maintenance), for discovering other bundles, and so forth. OSGi
bundles are particularly useful for controlling the functionality of networked appliances. Possible use cases include SetTopBoxes, Vehicles (note that car electronics today requires much shorter maintenance cycles than the mechanical parts, especially for software updates!), consumer electronics, and so forth. As to smart homes, the favored concept is that of a residential gateway that is connected to the global Internet and receives updates for smart home appliances via OSGi. The residential gateway may then forward bundle updates and so forth to the relevant appliances if needed.

OSGi has a number of deficiencies. For instance, it is not considered to be very resource effective. Nevertheless, it has tremendous impact as a de facto standard for dealing with some of the elementary aspects of coping with global UC systems in a platform and vendor independent way.

REFERENCE ARCHITECTURES FOR UBIQUITOUS COMPUTING

The Importance and Role of a Reference Architecture

A sophisticated distributed infrastructure is needed in order to make a myriad of networked UC nodes communicate and cooperate. If interoperability is to take on a worldwide scale, means for agreement among arbitrary participants must be provided. Ideally, the move from isolated proprietary UC solutions to a world of cooperating UC components is driven by so-called reference architectures which establish several levels of agreement: on level one, a common terminology and conceptualization of UC systems is established in order for researchers and practitioners to speak the same language and to work on the same global UC vision. On the second level, a common understanding of the ensemble and components of a typical UC system is established, including the potential roles of the components. On level three, basic functional principles can then be agreed upon. A fourth level is desirable but beyond the scope of reference architectures, that is concrete standards for intercomponent cooperation. This level is discussed in the introduction to the part Scalability.

Reference Architectures in a More Realistic World

In reality, a worldwide common understanding and corresponding standards have to be developed in a struggle for the best solution. Real life has a large impact on what becomes widespread. By “real life” we mean breaking research results, industry practice, experiences gained with proprietary prototypes and realizations, user acceptance, and not least business interests defended by global industrial players. Nevertheless, the exercise of proposing and refining reference architectures—in communication with the stakeholders mentioned—plays a key role in a struggle for globally interoperable solutions. Here reference architectures must be invented and published and then consolidated and reiterated based on feedback by the stakeholders.

Prominent Examples from the Past

The ISO reference architecture for open systems interconnection (OSI) was developed in the 1970s as an important step towards global networks. OSI was very successful in that it led to a common terminology and a common understanding of the components of computer networks including their roles. The fourth level aforementioned above: ISO standards for communication protocol, were not nearly as successful as the reference architecture itself. Rather, the Internet protocols TCP and IP took over almost the entire market. Nevertheless, the OSI reference architecture was extremely influential on the computer networking community as a whole and on the Internet in particular.
Another ISO reference architecture is ODP (open distributed processing). It emphasizes complex distributed systems and applications. An influential contribution of ODP is its support for different viewpoints of various stakeholders. In particular, ODP emphasized the importance of enterprise modeling for application development. All too often, applications are modeled and built with a technology focus and thus neglect the (dynamically changing) organization they should support. ODP addresses important issues, but came at a time when distributed applications were usually rather simple: ODP was considered overkill.

Layered Architectures vs. Component Architectures

Before we introduce concrete reference architectures, it is worth recalling the two complementary flavors:

- **Layered reference architectures** serve as a blueprint for layered software architectures. Both arrange sets of functions into layers that act as virtual machines: only the “what” (provided functionality and how to access it) must be known to users in higher layers, whereas the internal “how” (realization) is hidden and can be independently modified. The layer stack represents the range from higher to lower function sets, where higher means “closer to what users and applications need” and lower means “closer to what hardware provides”. Strict variants preclude higher layer components to access lower layers except for the one immediately below. Recent research has concentrated on approaches for automatic, selective custom configuration of the entire layer stack, according to the needs of applications—this trend is important in the UC world where dedicated, resource-poor UC nodes cannot host fat all-purpose layers.

- **Component reference architectures** take a birds-eye view on the world addressed. They define a number of cooperating components or rather component types, and specify inter-component cooperation at a certain level of detail. Again, a kind of art of right-sizing exists: too few component types do not really help to understand and discern relevant roles and specializations common to the world addressed, too many component types lead to overly complex architectures and problems in matching reference and reality.

Although we focus on the Computer Networks / Distributed Systems aspects of UC in the remainder of this chapter, readers should note that the entire book represents a holistic approach.

Why Component Reference Architectures are Important for UC

The OSI reference architecture assumes a network consisting of rather homogeneous nodes, namely general-purpose computers with ‘sufficient’ CPU and memory capacity. Accordingly, a common definition of a computer network reads as follows:

\[
\text{A computer network } CN \text{ is a set of autonomous nodes } AN, \text{ each of which disposes of CPU(s) and memory, plus a Communication Subsystem CSS capable of exchanging messages between any of the nodes: } CN := \{AN\} \cup CSS.
\]

In the definition, “all nodes are created equal”. At a closer look, computer networks rely on four mandatory constituents of nodes (ANs):

1. **Communication capability**: The capacity of exchanging messages with other nodes through the CSS.
2. **Address**: A unique identifier that can be used to specify the recipient or sender of messages.
3. **Processor**: A general purpose CPU.
4. **Memory**: Means for storing—at least—incoming messages.

In a UC world, resource scarcity and the special-purpose nature of many nodes are key issues.

A holistic UC approach must scale from servers to sensors and support the consideration of smart labels etc. The definition of a UC node must be different from the one above—the four constituents now read as follows:

1. **Communication** is mandatory, but may be passive (cf. passive RFID tags)
2. **Address** is not necessarily a unique identifier; for example, in a sensor network, a random node out of a redundant set with identical address may provide a certain functionality
3. **Processor** becomes an *optional* constituent
4. **Memory** becomes an *optional* constituent, too

With the above modifications, not all nodes are *autonomous (ANs)* any more.

**Proposed UC Component Reference Architectures**

The definition introduces a first possibility for distinguishing nodes as components of an application, that is, from the component architecture point of view. However, it only discerns between existing versus missing fundamental characteristics. More interesting is the aspect of different roles that nodes can play in the network—not application specific roles, but fundamental roles in the set of cooperating resources. Thus UC systems will take on more complex node topologies than what was considered in the eras of simple interprocess communication and client-server computing. In addition, a holistic approach needed for UC systems raises issues such as security, which are important when trying to find important node types at different levels of granularity.

One of the first proposals for a UC component reference architecture was made by the Fraunhofer research institute FOKUS in Berlin. They did not distinguish different node types that would assume different roles, but identified important roles that each UC node may potentially assume. Their concept is coined *I-Centric Services* and achieved a certain level of influence on the industrial Object Management Group (OMG). In their view, a UC node (usually a software service) should provide standard interfaces for four major issues:

1. **Discovery** of peers in a spontaneous, configuration-free manner
2. **Maintainance**, i.e., software update and revision
3. **Reservation**, that is, pre-allocation of some of the node’s resources as a basis for service guarantees
4. **Configuration** as a means for customizing the service for a dedicated role

Nodes that conform to these interfaces are called *super distributed objects (SDO)* in this proposal.

We will discuss another component architecture in some more detail since it attempts to discern between more specific roles of UC nodes. It was developed in the Telecooperation Group at the Technische Universität Darmstadt and is called *Mundo*, see Figure 4. *Mundo* distinguishes five different node types: *Me, Us, It, We, and They*.

*Me (Minimal Entity):* *Mundo* emphasizes the importance of a distinct personal UC node, that is the device tightly associated with its user: the Me. Every user uses exactly only one *Me* at any time. The rationale is rooted in the envisioned ubiquity of computer support in everyday life: if every step that one takes is potentially computer supported and controlled, then humans need a high level of trust that the computers “do the right thing”. For instance, users will want to make sure
that their actions are only recorded and disclosed to the degree they consent to or that is legally imposed. As another example, they want to be sure that they only trigger actions which they understand in their legal, financial, and other consequences and that they agree to. To this end, the Mundo researchers propose to conceptualize a single, truly owned UC node type that acts in the user’s stead and controls when, how, and to what extent other UC node types are invited or chartered to participate in actions. Since computer use becomes ubiquitous, such a personally-owned node type must be carried along virtually at all times. This imposes strong requirements with respect to miniaturization, robustness, and the conflicting goals of (a) the impossibility to falsify or duplicate such a node, and (b) the possibility to replace it easily in case of theft or failure. An important research questions is concerned with the minimum functionality of a Me.

Me nodes are considered as the representation of their users in the digital world—a digital persona involved in all user activities. It is a small wearable computer with minimal functionality. In order to support interaction with UC environments in a sensible way, the term minimal must be associated with a set of specific requirements regarding size, identity, security, interaction, context awareness, and networking. The design was guided by the principle that the minimal feature set of a system is determined by the worst-case environmental conditions under which the application must run satisfactorily (Satyanarayanan, 2001). This leads to a focus on speech based interaction and it is described in detail by Aitenbichler, Kangasharju, and Mühlhäuser (2004). Any Me can augment its capabilities through association with other entities of the Mundo architecture as described next.

Us (Ubiquitous aSSociable object): Minimization pressure will not permit feature-rich Mes. Hence, they must be able to connect to other mobile devices or devices embedded into the environment to offer more powerful services to their users, such as large display space. This process is called association and such devices are called ubiquitous aSSociable objects (Us). A Us is a computing device that extends the user’s personal environment by adding storage, processing capacity, displays, interaction devices, and so forth. During association, the Me sends authentication information to the Us, sets up a secure communication link, and personalizes the Us to suit the user’s preferences and needs. For privacy reasons, any personalization of a Us becomes automatically unavailable if it is out of range of the user’s Me.

Ir (smart iTem): There are also numerous smart items that do not support association that would classify them as Us. Vending machines, goods equipped with radio frequency IDs, and landmarks with “what is” functionality are just a few examples. Such devices are called smart items (Irs). An Ir is any digital or real entity that has an identity and can communicate with a Us or the Me. Communication may be active or passive. Memory and computation capabilities are optional (cf. the four constituents of a UC node described previously).

We (Wireless group Environment): Ad-hoc networking is restricted to an area near to the user of a Me device, as connections with remote services will involve a non ad hoc network infrastructure. The functionality of a wireless group environment is to bring together two or more personal environments consisting of a Me and arbitrary Us entities each. It enables cooperation between the devices and also allows for sharing
and transferring hardware (e.g., Us devices) and software or data between We users.

**THEY (Telecooperative Hierarchical ovErlaY)** stands for the backbone infrastructure as part of the Mundo component architecture. It connects users to the (nonlocal) world, and delivers services and information to the user. The THEY integrates different physical networks and provides transparent data access to users. Frequently used data may be cached on Us devices.

**UC Layered Reference Architectures**

Many actual UC projects are based on a layered architecture. Most of them are just first approaches to software architectures, only a few of them are intended to serve as a crystallization point for the community and future standards. Nevertheless, one of them may turn out to be so successful that a future reference architecture will evolve from it. We will concentrate on a small selection of the few projects that have a general reference model in mind. They concentrate on different challenges or foci, that is their findings will have to be merged if a holistic layered architecture is to be derived.

A first focus is the enterprise modeling that ODP already addressed. A reference architecture worth mentioning here is ODSI, the so-called open distributed services infrastructure (Bond, 2001). Although already outdated, ODSI was influential since it fostered the move away from ODP’s more top-down approach to a component-based, that is service based approach that supports the concept of applications being compositions of services.

Other reference architectures emphasize Smart Environments. Two facets are important and investigated—still—in different camps even as to the work on reference architectures: smart information spaces and smart physical spaces. By smart information spaces, we mean environments which concentrate on cooperative treatment of IT- and data/media centric work (cf. Mark Weiser’s three initial UC devices). Smart physical spaces are often called smart spaces or more specifically smart houses, labs, offices, homes etc. Work on these kinds of environments emphasizes the tangible, physical (computer-augmented) objects to be handled.

As for smart information spaces, an interesting reference architecture was proposed in the LifeSpaces project in South Australia (Bright & Vernik, 2004). Their architecture incorporates some of the findings from ODSI and distinguishes four layers:

1. **Enterprise model**: This layer supports rules, processes, and organizational models of roles and services in the enterprise.
2. **Coordination and control including interaction support**: On this layer, a shared and persistent event space of limited capacity, and an agent-based workspace infrastructure are offered.
3. **Enterprise bus**: This term refers to a communication layer based on the publish/subscribe paradigm.
4. **The service layer**: Here, the core functionality is represented by easily composable services. An enterprise bus is offered for services to communicate and cooperate; this bus connects so-called peers which host the services.

As for smart physical spaces, a prominent example is the reference architecture developed by the Gator Tech Smart House project of the University of Florida (see Figure 5). The reference architecture depicted is a more recent version of what was published by Helal, Mann, El-Zabadani, King, Kaddoura, and Jansen (2005) and is included by courtesy of the authors (the commercial version is called Atlas now). For more information, the reader may consult the group’s Web Site at [www.icta.ufl.edu](http://www.icta.ufl.edu) or the Atlas Web site at [www.pervasa.com](http://www.pervasa.com). The architecture emphasizes sensors (plus actuators) and networked embedded devices at the lowest layer as the hardware foundation of UC...
applications. The OSGI standard is exploited for customizing and maintaining these sensors and embedded devices in a dedicated second layer. The third layer contains three large parts which reflect major insights into the nature of the UC world (note that these insights have a large influence on the present book, too):

- **The context management layer** reflects the importance of context-awareness for UC as a whole, as discussed in the preface of the book;
- **The service layer** reflects services (and service-oriented architectures, SOA) as the dominating paradigm for building autonomous software components in a UC setting;
- **The knowledge layer** reflects the fact that large-scale service composition cannot rely on standardized interfaces that are distributed prior to software (service) development; rather, service discovery and service interaction must rely on machine readable descriptions of the service semantics available at runtime;
- Due to the strictly service-oriented concept used, application development boils down to service composition; the top layer offers corresponding tools.

In conclusion, it should have become clear that both a component based and a layered reference architecture, if widely accepted, would be important steps from UC islands towards truly global UC. The reference architectures presented could serve as a basis for better communication among the UC protagonists and for the necessary standards.

**REFERENCES**


**ADDITIONAL READING**


---

This work was previously published in Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises, edited by M. Mühlhäuser and I. Gurevych, pp. 1-20, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 1.2
Ubiquitous Computing and the Concept of Context

Antti Oulasvirta
Helsinki Institute for Information Technology, Finland

Antti Salovaara
Helsinki Institute for Information Technology, Finland

INTRODUCTION

Mark Weiser (1991) envisioned in the beginning of the 1990s that ubiquitous computing, intelligent small-scale technology embedded in the physical environment, would provide useful services in the everyday context of people without disturbing the natural flow of their activities.

From the technological point of view, this vision is based on recent advances in hardware and software technologies. Processors, memories, wireless networking, sensors, actuators, power, packing and integration, optoelectronics, and biomaterials have seen rapid increases in efficiency with simultaneous decreases in size. Moore’s law on capacity of microchips doubling every 18 months and growing an order of magnitude every five years has been more or less accurate for the last three decades. Similarly, fixed network transfer capacity grows an order of magnitude every three years, wireless network transfer capacity every 5 to 10 years, and mass storage every 3 years. Significant progress in power consumption is less likely, however. Innovations and breakthroughs in distributed operating environments, ad hoc networking, middleware, and platform technologies recently have begun to add to the ubiquitous computing vision on the software side.

Altogether, these technological advances have a potential to make technology fade into the background, into the woodwork and fabric of everyday life, and incorporate what Weiser (1991) called natural user interfaces. Awareness of situational factors (henceforth, the context) consequently was deemed necessary for this enterprise. This article looks at the history of the concept of context in ubiquitous computing and relates the conceptual advances to advances in envisioning human-computer interaction with ubiquitous computing.
BACKGROUND

Ubiquitous Computing Transforms Human-Computer Interaction

Human-computer interaction currently is shifting its focus from desktop-based interaction to interaction with ubiquitous computing beyond the desktop. Context-aware services and user interface adaptation are the two main application classes for context awareness. Many recent prototypes have demonstrated how context-aware devices could be used in homes, lecture halls, gardens, schools, city streets, cars, buses, trams, shops, malls, and so forth.

With the emergence of so many different ways of making use of situational data, the question of what context is and how it should be acted upon has received a lot of attention from researchers in HCI and computer science. The answer to this question, as will be argued later, has wide ramifications for the design of interaction and innovation of use purposes for ubiquitous computing.

HISTORY

Context as Location

In Weiser’s (1991) proposal, ubiquitous computing was realized through small computers distributed throughout the office. Tabs, pads, and boards helped office workers to access virtual information associated to physical places as well as to collaborate over disconnected locations and to share information using interfaces that take locational constraints sensitively into account. Although Weiser (1991) never intended to confine context to mean merely location, the following five years of research mostly focused on location-based adaptation. Want et al. (1992) described the ActiveBadge, a wearable badge for office workers that could be used to find and notify people in an office. Weiser (1993) continued by exploring systems for sharing drawings between disconnected places (the Tivoli system). Schilit et al. (1994) defined context to encompass more than location—to include people and resources as well—but their application examples were still mostly related to location sensing (i.e., proximate selection, location-triggered reconfiguration, location-triggered information, and location-triggered actions). Want, et al. (1995) added physical parameters like time and temperature to the definition. Perhaps the best-known mobile application developed during this location paradigm era was the CyberGuide (Long et al., 1996), an intelligent mobile guide that could be used to search for nearby services in a city. This paradigm was also influential in the research on Smart Spaces, such as intelligent meeting rooms.

The Representational Approach to Context

Although the idea that location equals context was eventually dismissed, many researchers coming from computer science still believed that contexts were something that should be recognized, labeled, and acted upon (Schmidt et al., 1998). Here, context was supposed to be recognized from sensor data, labeled, and given to applications that would use it as a basis for adaptation. Dey et al.’s (1999) five Ws of context—Who, Where, When, What, and Why—extended this approach and demonstrated convincing examples of how a labeled context could be used for presenting, executing, and tagging information. Tennenhouse’s (2000) proactive computing paradigm endorsed a similar way of thinking about context, emphasizing the role of computers in doing real-time decisions on behalf of (or pro) the user. A somewhat similar approach that also attempts to delegate decision-making responsibility to intelligent systems is taken by the Ambient Intelligence (AmI) technology program of the European Union (ISTAG). One part of the AmI vision entails intelligent agents that assume some of the control responsibility from the users.
The latest widely referred to definition was given by Dey et al. (2001), who defined context as “any information that characterizes a situation related to the interaction between users, applications, and the surrounding environment” (p. 106). Satyanarayanan’s (2001) formulation of pervasive computing also belongs to this line of thinking, but the author has chosen to avoid defining context and merely admits that it is rich and varies.

In his review of context definitions over the years, Dourish (2004) calls this the representational approach to context. Recent work within this branch has come close to finding the limits to recognizing and labeling contexts. For example, simple physical activities of a person in a home environment can be recognized with about 80-85% accuracy (Intille et al., 2004), as can be the interruptability of a person working in an office (Fogarty et al., 2004). Some critics have drawn parallels from this enterprise to problems encountered in strong AI (Erickson, 2002).

**FUTURE TRENDS**

**New Directions Inspired by Human and Social Sciences**

By the year 1996, other approaches to context were beginning to emerge. Wearable computing (Mann, 1996) looked at personal wearable computers able to help us remember and capture our everyday experiences through video and sound recording of context. Tangible bits (Ishii & Ullmer, 1997), although inspired by ubiquitous computing, looked at context not as something that had to be reacted upon but as surroundings of the user that could be augmented with tangible (i.e., graspable) computers and ambient media that display digital information using distraction-free output channels.

More recently, researchers have started to emphasize the social context and issues in people’s practices and everyday conduct. These approaches give special consideration to activities that people engage in and highlight their dynamic nature, different from the labeling-oriented representational approach. Activity-centered approaches emphasize both turn taking in communication between the user and the applications (Fischer, 2001) and acknowledge the situated and time-varying nature of the needs that a user has in his or her life (Greenberg, 2001). This line of research highlights the difficulties that exist in making correct inferences about a user’s tasks through sensor information. Considerations of social issues in ubiquitous computing design include questions of how to fit computation intelligence into people’s routines in an unremarkable manner (Tolmie et al., 2002) and how people’s patterns of interaction with humans and computers change when computationally augmented artifacts are adopted into use. Yet another emerging idea from HCI addresses specifically the aim to be free from distraction; that is, when it is appropriate to interrupt the user at his or her present task. Some call systems with such inferring capabilities *attentive user interfaces* (Vertegaal, 2003).

**CONCLUSION**

HCI in ubiquitous computing has been both inspired and constrained by conceptual developments regarding the concept of context. Weiser’s (1991) initial work caused researchers to conceive context narrowly as encompassing mainly location and other static, easily measurable features of a user’s context. After about five years of research, the restrictiveness of this definition was realized, and broader definitions were formulated. Still, context mainly was pursued by computer scientists and seen as something that must be labeled and reacted upon to adapt user interfaces. More recent work by human and social scientists has emphasized the role of user studies and theoretical reasoning in understanding what context entails in a particular application.
REFERENCES


**KEY TERMS**

**Attentive User Interfaces:** AUIs are based on the idea that modeling the deployment of user attention and task preferences is the key for minimizing the disruptive effects of interruptions. By monitoring the user’s physical proximity, body orientation, eye fixations, and the like, AUIs can determine what device, person, or task the user is attending to. Knowing the focus of attention makes it possible in some situations to avoid interrupting the users in tasks that are more important or time-critical than the interrupting one.

**Context:** That which surrounds and gives meaning to something else. (Source: The Free On-line Dictionary of Computing, http://foldoc.doc.ic.ac.uk/foldoc/)

**Peripheral Computing:** The interface attempts to provide attentionally peripheral awareness of people and events. Ambient channels provide a steady flow of auditory cues (i.e., a sound like rain) or gradually changing lighting conditions.

**Pervasive Computing:** Technology that provides easy access to information and other people anytime and anywhere through a mobile and scalable information access infrastructure.

**Proactive Computing:** A research agenda of developing interconnected devices and agents, equipped with faster-than-human-speed computing capabilities and means to affect real-world phenomena that a user can monitor and steer without a need to actively intervene in all decision-making situations. By raising the user above the traditional human-computer interaction loop, efficiency and freedom from distraction are expected to be enhanced.

**Tangible Bits:** According to Hiroshi Ishii of MIT, “the smooth transition of users’ focus of attention between background and foreground using ambient media and grasparable objects is a key challenge of Tangible Bits” (p. 235).

**Tangible User Interfaces:** Systems that give a physical form to digital information through augmenting tools and grasparable objects with computing capabilities, thus allowing for smooth transitions between the background and foreground of the user’s focus of attention.

**Unremarkable Computing:** An approach that focuses on designing domestic devices that are unremarkable to users. Here, unremarkable is understood as the use of a device being a part of a routine, because, it is believed, routines are invisible in use for those who are involved in them.

**Wearable Computing:** Technology that moves with a user and is able to track the user’s motions both in time and space, providing real-time information that can extend the user’s knowledge and perception of the environment.
INTRODUCTION

Although gender differences in a technological world are receiving significant research attention, much of the research and practice has aimed at how society and education can impact the successes and retention of female computer science professionals. The possibility of gender issues within software, however, has received almost no attention, nor has the population of female end users. However, there is relevant foundational research suggesting that gender-related factors within a software environment that supports end-user computing may have a strong impact on how effective male and female end users can be in that environment. Thus, in this article, we summarize theory-establishing results from other domains that point toward the formation of grounded hypotheses for studying gender differences in end-user computing.

There has been much background research relevant to human issues of end-user computing, which we define here as problem-solving using computer software, also termed end-user programming in some of the literature (e.g., Blackwell, 2002; Green & Petre, 1996; Nardi, 1993). (See the glossary for definitions of these and related terms.) Despite this, few researchers have considered potential gender HCI issues and gender differences that may need to be accounted for in designing end-user computing environments. The most notable exception is Czerwinski’s pioneering research on the support of both genders in navigating through 3-D environments (Czerwinski, Tan, & Robertson, 2002; Tan, Czerwinski, & Robertson, 2003). Although individual differences, such
as experience, cognitive style, and spatial ability, are likely to vary more than differences between gender groups, evidence from Czerwinski’s work as well as work in other domains, such as psychology and marketing, has found gender differences relevant to computer usage. In fact, some research has shown that some software is (unintentionally) designed for males (Huff, 2002).

One reason gender HCI issues in end-user computing are important is that ignorance of gender issues has already proven to be dangerous: today’s low percentage of computer science females (Camp, 1997) has been directly attributed to the past unawareness of gender issues in computer science education and in the workforce. There is a risk that if gender HCI issues in end-user computing environments are ignored, a similar phenomenon could occur with female end users.

**WHAT COULD GO WRONG?**

What gender differences might matter in the design of end-user computing environments? Consider the following scenario in one particular end-user computing environment.

Imagine a female teacher engaged in preparing a spreadsheet to track her students’ scores and to calculate ways of providing students with the best grades. Part of her process of preparing her spreadsheet is to test the spreadsheet. While she is engaged in testing, the system surprises her by decorating some of the spreadsheet cells, as in Figure 1.

The surprises were intentionally placed into the software by the designers relying on a strategy for end-user computing environments called Surprise-Explain-Reward (Wilson et al., 2003). The surprise, which was intended to capture the teacher’s attention and arouse her curiosity, reveals the presence of an “information gap” (Lowenstein, 1994). In this case the system is using the surprise to interest her in assertions (Burnett et al., 2003), which she can use to guard against future errors by specifying, for example, that the value of a cell calculating a grade average should always fall between 0 and 100.

What could go wrong in surprising the user? According to Lowenstein’s information gap theory, a user needs to have a certain level of confidence in order to reach a useful level of curiosity (Lowenstein, 1994). However, given documented gender differences in computer confidence (Busch, 1995; Huff, 2002), the teacher’s level of computer confidence could interfere with the surprise’s ability to capture her interest.

Returning to our scenario, suppose for this particular user, the surprise is effective at arousing her curiosity, she looks to the object that surprised her (the assertion) for an explanation. The explanation, viewed through a tooltip, includes the semantics, possible actions she can take (regarding the assertion), and the future reward(s) of taking the action. See Figure 1.

What could go wrong with the explanation? According to one theory, males and females process information differently (Meyers & Sternthal, 1991; O’Donnell & Johnson, 2001), and thus both the presentation and the content of the explanation may impact its effectiveness for males versus females. If the information needed by the user is not effectively communicated, the user’s ability to problem solve is likely to be reduced.

Another role of the explanation is to help users make a reasonably accurate assessment of the risk in taking some action—but since males and females differ in their perceptions of risk (Byrnes, Miller, & Schafer, 1999), the explanation may need to serve these two populations differently in this respect as well. (An example of risk may be the fear that the user will lose their work if they try a certain feature.) If one gender perceives an explanation of a feature as communicating higher levels of risk than another, the users with higher risk perceptions may avoid supposedly “risky” features that may be important to overall effectiveness.
Perhaps the most important role of explanations is to make clear the rewards of using particular features of the software. Providing information about rewards in the explanation is consistent with the implications of the Model of Attention Investment (Blackwell, 2002), an analytic model of user problem-solving behavior that models the costs, benefits, and risks users weigh in deciding how to complete a task. An implication of this model is that if the system provides the user an idea of future benefits, users can better assess if the cost of using a feature (here assertions) is worth their time. The reward aspect of the strategy refers to rewards such as the automatic detection of errors, which is depicted by the red circle around HomeWork1’s erroneous value in Figure 1.

What could go wrong with rewards? Since males and females are often motivated by different factors, there may be gender differences in what actually is a perceived “reward.” If the rewards are only tailored to one gender’s perceptions of rewards, the other gender may not be motivated to use the devices that will help them be effective. In this end-user computing scenario, potential problems arose that may be addressable within the end-user computing software itself. Four issues that arose here were (1) software features whose effects depend upon users’ computer confidence (discussed in the section on confidence), (2) the software’s ability to communicate effectively with users (discussed in support), (3) the possibility of a user’s perception of risk interfering with the user choosing to use appropriate features (discussed in motivation), and (4) possible differences between a user’s actual motivations and the software’s attempt to “reward” users for using particular features (discussed in motivation). These issues together form a useful organizational framework for considering gender HCI.

CONFIDENCE

This document uses the term “confidence” for the interrelated concepts of self-confidence, self-efficacy, overconfidence, and perceived risk.

From the field of computer science, there is substantial evidence of low confidence levels as computer science females compare themselves to the males (Margolis, Fisher, & Miller, 1999). Of particular pertinence to end-user computing, however, is the evidence showing that low confidence relating to technology is not confined to computer science females (Busch, 1995; Huff, 2002; Torkzadeh & Van, 2002).

As a measure of confidence, researchers often use self-efficacy, as was done in the Busch study. Self-efficacy is belief in one’s capabilities to perform a certain task (Bandura, 1994). There is specific evidence that low self-efficacy impacts
attitudes toward a new software package prior to its use (Hartzel, 2003). Taken together, this research suggests that a first experience with end-user computing software can impact attitudes which may in turn impact users’ future choices to (or not to) use some features in their software.

Overconfidence matters too, because it can prevent people from suspecting errors, leading to misplaced trust in erroneous programs. In particular, overconfidence in spreadsheet correctness is common (Panko, 1998). There is evidence (Lunderberg, Fox, & Punchochar, 1994) suggesting gender differences in overconfidence just as in under confidence. Hence, designing methods to help alleviate overconfidence in end-user computing needs to be carefully targeted specifically toward overconfident users.

Perception of risk is tied to confidence, and impacts the decisions people make. According to the attention investment model (a model of how users allocate their attention in problem solving) (Blackwell, 2002), a user may choose not to follow through a particular action, if they decide that the costs and/or risks are too high in relation to the benefits of taking that action. Perception of risk thus plays an important role in a user’s decision-making about whether to use some features, and can lead to differences in actual behavior.

Differences in perception of risk have been tied to gender. For example, a meta-analysis of 150 studies on gender and risk taking found that females engaged in less risk taking than males (Byrnes, Miller, & Schafer 1999). The meta-analysis did not address risks of computer use directly. However, it did find that intellectual risk taking, defined as activities involving mathematical or spatial reasoning skills, was greater in males than in females (Byrnes, Miller, & Schafer 1999).

To obtain evidence about whether confidence might directly impact gender HCI issues for end-user computing environments, we conducted a preliminary think-aloud study in which participants attempted to debug two spreadsheets, given the support of Surprise-Explain-Reward devices (Beckwith & Burnett, 2004). To our surprise, the females’ confidence levels dropped over the course of the study much more than did the males’ confidence levels. This result suggests that end-user computing environment itself—which, like other end-user computing environments, was designed without knowledge of gender-related HCI principles—is not currently serving the females’ needs as well as the males’. A subsequent follow-up study (Beckwith et al., 2005) confirmed this: in the spreadsheet paradigm, ties were found between females’ confidence issues and low utilization of features aimed to support problem solving, resulting in effectiveness problems.

**SUPPORT**

We will use the term “support” to mean built-in aspects of the software, such as on-line help systems and Figure 1’s tooltips that help users learn or understand the environment.

The system’s approach to help users achieve mastery in remembering the software’s devices may depend on a user’s learning style. One survey of university students found that students with an “abstract random” learning style were significantly more likely to be female and, as a result, could find computer-based instruction ineffective for learning (Ames, 2003). Other researchers have also found gender differences in learning styles (Heffler, 2001; Severiens & ten Dam, 1997). One implication of these findings is that end-user computing may need to support several learning styles, especially if some users are easily dissuaded by support devices not sensitive to their learning style.

Problem-solving style also shows gender differences, at least for computer games (Kafai, 1998). Researchers found that, unlike boys, rather than working in a linear fashion through the game, girls prefer to explore and move freely about a game (Gorriz & Medina, 2000). In another difference in problem-solving style, boys’ games
Gender and End-User Computing

typically depend upon competition, whereas girls prefer collaboration and working together (Kafai, 1998). For end-user computing environments, these problem-solving differences suggest differences in the support provided by the system. For example, supporting both linear and non-linear problem-solving styles and providing avenues for both competition and collaboration may be important for software’s success at supporting both genders adequately.

Finally, the theory of selectivity suggests gender differences which may impact how users process the information support devices provide. The theory of selectivity, from research in the area of marketing, states that males and females differ in their information processing strategies (Meyers & Sternthal, 1991; O’Donnell & Johnson, 2001). According to this theory of selectivity, females are more likely to employ elaborative information processing strategies, regardless of whether the task is simple or complex in nature. Males, however, are more likely to select heuristic processing strategies that minimize cognitive effort and reduce information load for simple tasks, switching to an elaborative strategy only on more complex tasks.

These gender differences have been shown to impact diverse software-related activities, ranging from users’ perceptions of Web sites used for e-commerce (Simon, 2001) to users’ performance on auditing tasks (O’Donnell & Johnson, 2001). For end-user computing environments, this research may have implications for informing end users of important information via the software’s support devices.

MOTIVATION

Research has shown that computer science females are motivated by how technology can help other people, whereas males tend to enjoy technology for its own sake (Margolis, Fisher, & Miller, 1999). These differences are also found with other females who use technology, such as architects, NASA scientists, and filmmakers. One study (Brunner, Bennett, & Honey, 1998) found that females described technological objects as tools to help integrate personal and professional lives and to facilitate creativity and communication, whereas males described them as technological devices to increase command and control over nature and one another. The gender differences found in that study are summarized in Table 1.

The technology acceptance model (TAM) (Morris & Dillon, 1997; Venkatesh & Morris, 2000) provides a model of users’ acceptance and usage behavior of technology. According to TAM, user acceptance, and ultimately technology use, is determined by two key beliefs: perceived usefulness and perceived ease of use (Venkatesh & Morris, 2000). “Perceived usefulness” is the degree to which a user believes that using the system will enhance their performance, and “perceived ease of use” is the degree to which the user believes that using the system will be free of effort. According to one study, the relative importance of each differs by gender (Venkatesh & Morris, 2000); women were more influenced by perceived ease of use whereas men were more influenced by perceived usefulness.

FUTURE TRENDS AND CONCLUSION

To date, there has been little research on how the design of software itself may interact with gender differences. Still, foundational work from several domains strongly suggests that such differences may have critical impacts on users’ success in end-user computing.

Although research is beginning to emerge providing some insights into gender’s importance in end-user computing environments, it is still largely an open question. Also open are questions of what specific types of differences matter in such environments and what amelioration strategies are possible.
Gender and End-User Computing

To help provide a foundation upon which researchers interested in these issues can build, this article has drawn from five domains to summarize literature relevant to these questions. All of the literature surveyed identifies one or more issues that potentially impact end users’ success that are also potentially addressable within the software system. Together, the open questions and survey are intended to provide a foundation for future investigation.

ACKNOWLEDGMENT

This work was supported in part by Microsoft Research and by the EUSES Consortium via NSF grants ITR 0325273 and CNS 0420533.

Table 1. Summary of gender differences in fantasizing about technology (Brunner, Bennett, & Honey, 1998). Reprinted with permission.

<table>
<thead>
<tr>
<th>Women ...</th>
<th>Men ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 fantase about it as a MEDIUM</td>
<td>fantasize about it as a PRODUCT</td>
</tr>
<tr>
<td>2 see it as a TOOL</td>
<td>see it as a WEAPON</td>
</tr>
<tr>
<td>3 want to use it for COMMUNICATION</td>
<td>want to use it for CONTROL</td>
</tr>
<tr>
<td>4 are impressed with its potential for CREATION</td>
<td>are impressed with its potential for POWER</td>
</tr>
<tr>
<td>5 see it as EXPRESSIVE</td>
<td>see it as a INSTRUMENTAL</td>
</tr>
<tr>
<td>6 ask it for FLEXIBILITY</td>
<td>ask it for SPEED</td>
</tr>
<tr>
<td>7 are concerned with its EFFECTIVENESS</td>
<td>are concerned with its EFFICIENCY</td>
</tr>
<tr>
<td>8 like its ability to facilitate SHARING</td>
<td>like its ability to facilitate AUTONOMY</td>
</tr>
<tr>
<td>9 are concerned with INTEGRATING it into their personal lives</td>
<td>are interested in CONSUMING it</td>
</tr>
<tr>
<td>10 talk about wanting to EXPLORE worlds</td>
<td>talk about using it to EXPLOIT resources and potentialities</td>
</tr>
<tr>
<td>11 are EMPOWERED by it</td>
<td>want TRANSCENDENCE</td>
</tr>
</tbody>
</table>

REFERENCES


O’Donnell, E., & Johnson, E. (2001). Gender effects on processing effort during analytical
End User: Users who are not trained programmers.

End-User Computing: Computer-supported problem solving by end users, using systems such as spreadsheets, multimedia authoring tools, and graphical languages for demonstrating the desired behavior.

End-User Programming: A term synonymous with end-user computing.

Gender HCI: Human-computer interaction (HCI) work that takes gender differences into account.

Overconfidence: Higher self-efficacy than is warranted by a user’s abilities.

Self-Efficacy: Belief in one’s capabilities to perform a certain task.

Under Confidence: Lower self-efficacy than is warranted by a user’s abilities.

This work was previously published in Encyclopedia of Gender and Information Technology, edited by E. Trauth, pp. 398-404, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global)
Chapter 1.4
The Theoretical Framework of Cognitive Informatics

Yingxu Wang
University of Calgary, Canada

ABSTRACT

Cognitive Informatics (CI) is a transdisciplinary enquiry of the internal information processing mechanisms and processes of the brain and natural intelligence shared by almost all science and engineering disciplines. This article presents an intensive review of the new field of CI. The structure of the theoretical framework of CI is described encompassing the Layered Reference Model of the Brain (LRMB), the OAR model of information representation, Natural Intelligence (NI) vs. Artificial Intelligence (AI), Autonomic Computing (AC) vs. imperative computing, CI laws of software, the mechanism of human perception processes, the cognitive processes of formal inferences, and the formal knowledge system. Three types of new structures of mathematics, Concept Algebra (CA), Real-Time Process Algebra (RTPA), and System Algebra (SA), are created to enable rigorous treatment of cognitive processes of the brain as well as knowledge representation and manipulation in a formal and coherent framework. A wide range of applications of CI in cognitive psychology, computing, knowledge engineering, and software engineering has been identified and discussed.

INTRODUCTION

The development of classical and contemporary informatics, the cross fertilization between computer science, systems science, cybernetics, computer/software engineering, cognitive science, knowledge engineering, and neuropsychology, has led to an entire range of an extremely interesting and new research field known as Cognitive Informatics (Wang, 2002a, 2003a, b, 2006b; Wang, Johnston & Smith 2002; Wang & Kinsner, 2006). Informatics is the science of information that studies the nature of information; it’s processing, and ways of transformation between information, matter, and energy.

Definition 1. Cognitive Informatics (CI) is a transdisciplinary enquiry of cognitive and information sciences that investigates the internal in-
formation processing mechanisms and processes of the brain and natural intelligence, and their engineering applications via an interdisciplinary approach.

In many disciplines of human knowledge, almost all of the hard problems yet to be solved share a common root in the understanding of the mechanisms of natural intelligence and the cognitive processes of the brain. Therefore, CI is a discipline that forges links between a number of natural science and life science disciplines with informatics and computing science.

The structure of the theoretical framework of CI is described in Figure 1, which covers the Information-Matter-Energy (IME) model (Wang, 2003b), the Layered Reference Model of the Brain (LRMB) (Wang, Wang, Patel & Patel, 2006), the Object-Attribute-Relation (OAR) model of information representation in the brain (Wang, 2006h; Wang & Wang, 2006), the cognitive informatics model of the brain (Wang, Liu, & Wang, 2003; Wang & Wang, 2006), Natural Intelligence (NI) (Wang, 2003b), Autonomic Computing (AC) (Wang, 2004), Neural Informatics (NeI) (Wang, 2002a, 2003b, 2006b), CI laws of software (Wang, 2006f), the mechanisms of human perception processes (Wang, 2005a), the cognitive processes of formal inferences (Wang, 2005c), and the formal knowledge system (Wang, 2006g).

In this article, the theoretical framework of CI is explained in the fundamental theories of CI section. Three structures of new descriptive
mathematics such as Concept Algebra (CA), Real-Time Process Algebra (RTPA), and System Algebra (SA) are introduced in the denotational mathematics for CI in order to rigorously deal with knowledge and cognitive information representation and manipulation in a formal and coherent framework. Applications of CI are discussed, which covers cognitive computing, knowledge engineering, and software engineering. Then, it draws conclusions on the theories of CI, the contemporary mathematics for CI, and their applications.

THE FUNDAMENTAL THEORIES OF CI

The fundamental theories of CI encompass 10 transdisciplinary areas and fundamental models, T1 through T10, as identified in Figure 1. This section presents an intensive review of the theories developed in CI, which form a foundation for exploring the natural intelligence and their applications in brain science, neural informatics, computing, knowledge engineering, and software engineering.

The Information-Matter-Energy Model

Information is recognized as the third essence of the natural world supplementing to matter and energy (Wang, 2003b), because the primary function of the human brain is information processing.

Theorem 1. A generic worldview, the IME model states that the natural world (NW) that forms the context of human beings is a dual world: one aspect of it is the physical or the concrete world (PW), and the other is the abstract or the perceptive world (AW), where matter (M) and energy (E) are used to model the former, and information (I) to the latter, that is:

\[
\text{NW} \equiv \text{PW} \parallel \text{AW} = p(M, E)\parallel a(I) = n(I, M, E)
\]  

where \(\parallel\) denotes a parallel relation, and \(p, a,\) and \(n,\) are functions that determine a certain \(PW, AW,\) or \(NW,\) respectively, as illustrated in Figure 2.

According to the IME model, information plays a vital role in connecting the physical world with the abstract world. Models of the natural world have been well studied in physics and other natural sciences. However, the modeling of the abstract world is still a fundamental issue yet to be explored in cognitive informatics, computing, software science, cognitive science, brain sciences, and knowledge engineering. Especially the relationships between I-M-E and their transformations are deemed as one of the fundamental questions in CI.

Corollary 1. The natural world \(NW(I, M, E),\) particularly part of the abstract world, \(AW(I),\) is cognized and perceived differently by individuals because of the uniqueness of perceptions and mental contexts among people.

Corollary 1 indicates that although the physical world \(PW(M, E)\) is the same to everybody, the natural world \(NW(I, M, E)\) is unique to different individuals because the abstract world \(AW(I),\) as a part of it, is subjective depending on the information an individual obtains and perceives.

Corollary 2. The principle of transformability between IME states that, according to the IME model, the three essences of the world are predicated to be transformable between each other as described by the following generic functions \(f_1\) to \(f_6:\)

\[
I = f_1(M)
\]

\[
M = f_2(I) = f_1^{-1}(I)
\]

\[
I = f_3(E)
\]

\[
NW = PW || AW = p(M, E) || a(I) = n(I, M, E)
\] (1)

\[
I = f_1(M)
\]

\[
M = f_2(I) = f_1^{-1}(I)
\]

\[
I = f_3(E)
\]
\[ E = f_4(I) \cdot f_3^{-1}(I) \quad (2.4) \]
\[ E = f_5(M) \quad (2.5) \]
\[ M = f_6(E) \cdot f_5^{-1}(E) \quad (2.6) \]

where a question mark on the equal sign denotes an uncertainty if there exists such a reverse function (Wang, 2003b).

Albert Einstein revealed Functions \( f_5 \) and \( f_6 \), the relationship between matter (m) and energy (E), in the form \( E = mC^2 \), where \( C \) is the speed of light. It is a great curiosity to explore what the remaining relationships and forms of transformation between I-M-E will be. To a certain extent, cognitive informatics is the science to seek possible solutions for \( f_1 \) to \( f_4 \). A clue to explore the relations and transformability is believed in the understanding of the natural intelligence and its information processing mechanisms in CI.

**Definition 2.** Information in CI is defined as a generic abstract model of properties or attributes of the natural world that can be distinctly elicited, generally abstracted, quantitatively represented, and mentally processed.

**Definition 3.** The measurement of information, \( I_k \), is defined by the cost of code to abstractly represent a given size of internal message \( X \) in the brain in a digital system based on \( k \), that is:

\[
I_k = f : X \rightarrow S_k = \lfloor \log_k X \rfloor \tag{3}
\]

where \( I_k \) is the content of information in a \( k \)-based digital system, and \( S_k \) is the measurement scale based on \( k \). The unit of \( I_k \) is the number of \( k \)-based digits (Wang, 2003b).

Equation 3 is a generic measure of information sizes. When a binary digital representation system is adopted, that is \( k = b = 2 \), it becomes the most practical one as follows.

**Definition 4.** The metalevel representation of information, \( I_b \), is that when \( k = b = 2 \), that is:

---

**Figure 3. LRMB model**

![LRMB model](image-url)
\[ I_b = f : X \rightarrow S_b \]
\[ = [\log_2 X] \]  

where the unit of information, \( I_b \), is a bit.

Note that the bit here is a concrete and deterministic unit, and it is no longer probability-based as in conventional information theories (Bell, 1953; Shannon, 1948). To a certain extent, computer science and engineering is a branch of modern informatics that studies machine representation and processing of external information; while CI is a branch of contemporary informatics that studies internal information representation and processing in the brain.

Theorem 2. The most fundamental form of information that can be represented and processed is binary digit where \( k = b = 2 \).

Theorem 2 indicates that any form of information in the physical (natural) and abstract (mental) worlds can be unified on the basis of binary data. This is the CI foundation of modern digital computers and NI.

The Layered Reference Model of the Brain

The LRMB (Wang et al., 2006) is developed to explain the fundamental cognitive mechanisms and processes of natural intelligence. Because a variety of life functions and cognitive processes have been identified in CI, psychology, cognitive science, brain science, and neurophilosophy, there is a need to organize all the recurrent cognitive processes in an integrated and coherent framework. The LRMB model explains the functional mechanisms and cognitive processes of natural intelligence that encompasses 37 cognitive processes at six layers known as the sensation, memory, perception, action, metacognitive, and higher cognitive layers from the bottom-up as shown in Figure 3. LRMB elicits the core and highly repetitive recurrent cognitive processes from a huge variety of life functions, which may shed light on the study of the fundamental mechanisms and interactions of complicated mental processes, particularly the relationships and interactions between the inherited and the acquired life functions as well as those of the subconscious and conscious cognitive processes.

The OAR Model of Information Representation in the Brain

Investigation into the cognitive models of information and knowledge representation in the brain is perceived to be one of the fundamental research areas that help to unveil the mechanisms of the brain. The Object-Attribute-Relation (OAR) model (Wang, 2006h; Wang et al., 2003) describes human memory, particularly the long-term memory, by using the relational metaphor, rather than the traditional container metaphor that used to be adopted in psychology, computing, and information science. The OAR model shows that human memory and knowledge are represented by relations, that is, connections of synapses between neurons, rather than by the neurons themselves as the traditional container metaphor described. The OAR model can be used to explain a wide range of human information processing mechanisms and cognitive processes.

The Cognitive Informatics Model of the Brain

The human brain and its information processing mechanisms are centred in CI. A cognitive informatics model of the brain is proposed in Wang and Wang (2006), which explains the natural intelligence via interactions between the inherent (subconscious) and acquired (conscious) life functions. The model demonstrates that memory is the foundation for any natural intelligence. Formalism in forms of mathematics, logic, and rigorous treatment is introduced into the study of cognitive and neural psychology and natural
informatics. Fundamental cognitive mechanisms of the brain, such as the architecture of the thinking engine, internal knowledge representation, long-term memory establishment, and roles of sleep in long-term memory development have been investigated (Wang & Wang, 2006).

**Natural Intelligence (NI)**

Natural Intelligence (NI) is the domain of CI. Software and computer systems are recognized as a subset of intelligent behaviors of human beings described by programmed instructive information (Wang, 2003b; Wang & Kinsner, 2006). The relationship between Artificial Intelligence (AI) and NI can be described by the following theorem.

**Theorem 3.** The law of compatible intelligent capability states that artificial intelligence (AI) is always a subset of the natural intelligence (NI), that is:

\[ AI \subseteq NI \]  

Theorem 3 indicates that AI is dominated by NI. Therefore, one should not expect a computer or a software system to solve a problem where humans cannot. In other words, no AI or computing system may be designed and/or implemented for a given problem where there is no solution being known by human beings.

**Neural Informatics (NeI)**

**Definition 5.** Neural Informatics (NeI) is a new interdisciplinary enquiry of the biological and physiological representation of information and knowledge in the brain at the neuron level and their abstract mathematical models (Wang, 2004; Wang & Wang, 2006).

NeI is a branch of CI, where memory is recognized as the foundation and platform of any natural or artificial intelligence (Wang & Wang, 2006).

**Definition 6.** The Cognitive Models of Memory (CMM) states that the architecture of human memory is parallel configured by the Sensory Buffer Memory (SBM), Short-Term Memory (STM), Long-Term Memory (LTM), and Action-Buffer Memory (ABM), that is:

\[ CMM = SBM \parallel STM \parallel LTM \parallel ABM \]  

where the ABM is newly identified in Wang and Wang (2006).

The major organ that accommodates memories in the brain is the cerebrum or the cerebral cortex. In particular, the association and premotor cortex in the frontal lobe, the temporal lobe, sensory cortex in the frontal lobe, visual cortex in the occipital lobe, primary motor cortex in the frontal lobe, supplementary motor area in the frontal lobe, and procedural memory in cerebellum (Wang & Wang, 2006).

The CMM model and the mapping of the four types of human memory onto the physiological organs in the brain reveal a set of fundamental mechanisms of NeI. The OAR model of information/knowledge representation described in the OAR model of information representation in the brain section provides a generic description of information/knowledge representation in the brain (Wang, 2006h; Wang et al., 2003).

The theories of CI and NeI explain a number of important questions in the study of NI. Enlightening conclusions derived in CI and NeI are such as: (a) LTM establishment is a subconscious process; (b) The long-term memory is established during sleeping; (c) The major mechanism for LTM establishment is by sleeping; (d) The general acquisition cycle of LTM is equal to or longer than 24 hours; (e) The mechanism of LTM establishment is to update the entire memory of information represented as an OAR model in the brain; and (f) Eye movement and dreams play an important
role in LTM creation. The latest development in CI and NeI has led to the determination of the magnificent and expected capacity of human memory as described in the Estimation of the Capacity of Human Memory section.

**Cognitive Informatics Laws of Software**

It is commonly conceived that software as an artifact of human creativity is not constrained by the laws and principles discovered in the physical world. However, it is unknown what constrains software. The new informatics metaphor proposed by the author in CI perceives software is a type of instructive and behavioral information. Based on this, it is asserted that software obeys the laws of informatics. A comprehensive set of 19 CI laws for software have been established in Wang (2006f), such as:

1. Abstraction
2. Generality
3. Cumulativeness
4. Dependency on cognition
5. Three-dimensional behavior space known as the object (O), space (S), and time (T)
6. Sharability
7. Dimensionless
8. Weightless
9. Transformability between I-M-E
10. Multiple representation forms
11. Multiple carrying media
12. Multiple transmission forms
13. Dependency on media
14. Dependency on energy
15. Wearless and time dependency
16. Conservation of entropy
17. Quality attributes of informatics
18. Susceptible to distortion
19. Scarcity

The informatics laws of software extend the knowledge on the fundamental laws and properties of software where the conventional product metaphor could not explain. Therefore, CI forms one of the foundations of software engineering and computing science.

**Mechanisms of Human Perception Processes**

**Definition 7.** *Perception* is a set of interpretive cognitive processes of the brain at the subconscious cognitive function layers that detects, relates, interprets, and searches internal cognitive information in the mind.

Perception may be considered as the *sixth sense* of human beings, which almost all cognitive life functions rely on. Perception is also an important cognitive function at the subconscious layers that determines personality. In other words, personality is a faculty of all subconscious life functions and experience cumulated via conscious life functions.

According to LRMB, the main cognitive processes at the perception layer are emotion, motivation, and attitude (Wang, 2005a). The relationship between the internal emotion, motivation, attitude, and the embodied external behaviors can be formally and quantitatively described by the *motivation/attitude-driven behavioral* (MADB) *model* (Wang & Wang, 2006), which demonstrates that complicated psychological and cognitive mental processes may be formally modeled and rigorously described by mathematical means (Wang, 2002b, 2003d, 2005c).

**The Cognitive Processes of Formal Inferences**

*Theoretical research* is predominately an inductive process, while *applied research* is mainly a deductive one. Both inference processes are based on the cognitive process and means of abstraction. *Abstraction* is a powerful means of philosophy and mathematics. It is also a preeminent trait of the human brain identified in CI studies (Wang,
All formal logical inferences and reasonings can only be carried out on the basis of abstract properties shared by a given set of objects under study.

**Definition 8.** *Abstraction* is a process to elicit a subset of objects that shares a common property from a given set of objects and to use the property to identify and distinguish the subset from the whole in order to facilitate reasoning.

Abstraction is a gifted capability of human beings. Abstraction is a basic cognitive process of the brain at the metacognitive layer according to LRMB (Wang et al., 2006). Only by abstraction can important theorems and laws about the objects under study be elicited and discovered from a great variety of phenomena and empirical observations in an area of inquiry.

**Definition 9.** *Inferences* are a formal cognitive process that reasons a possible causality from given premises based on known causal relations between a pair of cause and effect proven true by empirical arguments, theoretical inferences, or statistical regulations.

Formal inferences may be classified into the deductive, inductive, abductive, and analogical categories (Wang, 2005c). *Deduction* is a cognitive process by which a specific conclusion necessarily follows from a set of general premises. *Induction* is a cognitive process by which a general conclusion is drawn from a set of specific premises based on three designated samples in reasoning or experimental evidences. *Abduction* is a cognitive process by which an inference to the best explanation or most likely reason of an observation or event. *Analogy* is a cognitive process by which an inference about the similarity of the same relations holds between different domains or systems, and/or examines that if two things agree in certain respects, then they probably agree in others. A summary of the formal definitions of the five inference techniques is shown in Table 1.

For seeking generality and universal truth, either the objects or the relations can only be abstractly described and rigorously inferred by abstract models rather than real-world details.

**The Formal Knowledge System**

Mathematical thoughts (Jordan & Smith, 1997) provide a successful paradigm to organize and validate human knowledge, where once a truth or a theorem is established, it is true until the axioms or conditions that it stands for are changed or extended. A proven truth or theorem in mathematics does not need to be argued each time one uses it. This is the advantage and efficiency of formal knowledge in science and engineering. In other words, if any theory or conclusion may be argued from time-to-time based on a wiser idea or a trade-off, it is an empirical result rather than a formal one.

The Framework of Formal Knowledge (FFK) of mankind (Wang, 2006g) can be described as shown in Figure 5. An FFK is centered by a set of theories. A *theory* is a statement of how and why certain objects, facts, or truths are related. All objects in nature and their relations are constrained by invariable laws, no matter if one observed them or not at any given time. An *empirical truth* is a truth based on or verifiable by observation, experiment, or experience. A *theoretical proposition* is an assertion based on formal theories or logical reasoning. Theoretical knowledge is a formalization of generic truth and proven abstracted empirical knowledge. Theoretical knowledge may be easier to acquire when it exists. However, empirical knowledge is very difficult to be gained without hands-on practice.

According to the FFK model, an immature discipline of science and engineering is characterized by its body of knowledge not being formalized. Instead of a set of proven theories, the immature disciplines document a large set of
observed facts, phenomena, and their possible or partially working explanations and hypotheses. In such disciplines, researchers and practitioners might be able to argue every informal conclusion documented in natural languages from time-to-time probably for hundreds of years, until it is formally described in mathematical forms and proved rigorously.

The disciplines of mathematics and physics are successful paradigms that adopt the FFK formal knowledge system. The key advantages of the formal knowledge system are its stability and efficiency. The former is a property of the formal knowledge that once it is established and formally proved, users who refers to it will no longer need to reexamine or reprove it. The latter is a property of formal knowledge that is exclusively true or false that saves everybody’s time from arguing a proven theory.

**DENOTATIONAL MATHEMATICS FOR CI**

The history of sciences and engineering shows that new problems require new forms of mathematics. CI is a new discipline, and the problems in it require new mathematical means that are descriptive and precise in expressing and denoting human and system actions and behaviors. Conventional analytic mathematics are unable to solve the fundamental problems inherited in CI and related disciplines such as neuroscience, psychology, philosophy, computing, software engineering, and knowledge engineering. Therefore, denotational mathematical structures and means (Wang, 2006c) beyond mathematical logic are yet to be sought.

Although there are various ways to express facts, objects, notions, relations, actions, and behaviors in natural languages, it is found in CI that human and system behaviors may be classified into three basic categories known as to be, to have, and to do. All mathematical means and forms, in general, are an abstract and formal description of these three categories of expressibility and their rules. Taking this view, mathematical logic may be perceived as the abstract means for describing “to be,” set theory describing “to have,” and algebras, particularly process algebra, describing “to do.”

**Table 1. Definitions of formal inferences**

<table>
<thead>
<tr>
<th>Inference technique</th>
<th>Formal description</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Abstraction (\forall S, p \Rightarrow \exists e \in E \subseteq S, p(e))</td>
<td>-</td>
<td>To elicit a subset of elements with a given generic property.</td>
</tr>
<tr>
<td>2. Deduction (\forall x \in X, p(x) \Rightarrow \exists a \in X, p(a))</td>
<td>(\forall x \in X, p(x) \Rightarrow q(x)) \land (\exists a \in X, p(a) \Rightarrow q(a))</td>
<td>To derive a conclusion based on a known and generic premises.</td>
</tr>
<tr>
<td>3. Induction ((\exists a \in X, p(a) \land (\exists k, k+1 \in X, ((p(k) \Rightarrow p(k+1)))) \Rightarrow \forall x \in X, p(x))</td>
<td>((\exists a \in X, p(a) \Rightarrow q(a)) \land (\exists k, k+1 \in X, ((p(k) \Rightarrow q(k)) \Rightarrow (p(k+1) \Rightarrow q(k+1)))) \Rightarrow \forall x \in X, p(x) \Rightarrow q(x))</td>
<td>To determine the generic behavior of the given list or sequence of recurring patterns by three samples.</td>
</tr>
<tr>
<td>4. Abduction ((\forall x \in X, p(x) \Rightarrow q(x)) \land (\exists a \in X, q(a) \Rightarrow p(a)))</td>
<td>(\forall x \in X, p(x) \Rightarrow (q(x) \land r(x) \Rightarrow q(x))) \land (\exists a \in X, q(a) \Rightarrow (p(a) \lor r(a)))</td>
<td>To seek the most likely cause(s) and reason(s) of an observed phenomenon.</td>
</tr>
<tr>
<td>5. Analogy (\exists a \in X, p(a) \Rightarrow \exists b \in X, p(b))</td>
<td>(\exists a \in X, p(a) \Rightarrow q(a)) \land (\exists b \in X, p(b) \Rightarrow q(b))</td>
<td>To predict a similar phenomenon or consequence based on a known observation.</td>
</tr>
</tbody>
</table>
The Theoretical Framework of Cognitive Informatics

Theorem 4. The utility of mathematics is the means and rules to express thought rigorously and generically at a higher level of abstraction.

Three types of new mathematics, Concept Algebra (CA), Real-Time Process Algebra (RTPA), and System Algebra (SA), are created in CI to enable rigorous treatment of knowledge representation and manipulation in a formal and coherent framework. The three new structures of contemporary mathematics have extended the abstract objects under study in mathematics from basic mathematical entities of numbers and sets to a higher level, that is, concepts, behavioral processes, and systems. A wide range of applications of the denotational mathematics in the context of CI has been identified (Wang, 2002b, 2006d, e).

Concept Algebra

A concept is a cognitive unit (Ganter & Wille, 1999; Quillian, 1968; Wang, 2006e) by which the meanings and semantics of a real-world or an abstract entity may be represented and embodied based on the OAR model.

Definition 10. An abstract concept \( c \) is a 5-tuple, that is:

\[
c \triangleq (O, A, R^e, R^r, R^\prime)
\]  

where

- \( O \) is a nonempty set of object of the concept, \( O = \{o_1, o_2, ..., o_m\} \subseteq U \), where \( \mathcal{P}U \) denotes a power set of \( U \).
- \( A \) is a nonempty set of attributes, \( A = \{a_1, a_2, ..., a_n\} \subseteq \mathcal{P}M \).
• \( R^c \subseteq O \times A \) is a set of internal relations.
• \( R^i \subseteq C' \times C \) is a set of input relations, where \( C' \) is a set of external concepts.
• \( R^o \subseteq C \times C' \) is a set of output relations.

A structural concept model of \( c = (O, A, R^c, R^i, R^o) \) can be illustrated in Figure 6, where \( c, A, O, \) and \( R, R = \{R^c, R^i, R^o\} \), denote the concept, its attributes, objects, and internal/external relations, respectively.

**Definition 11.** Concept algebra is a new mathematical structure for the formal treatment of abstract concepts and their algebraic relations, operations, and associative rules for composing complex concepts and knowledge (Wang, 2006e).

Concept algebra deals with the algebraic relations and associational rules of abstract concepts. The associations of concepts form a foundation to denote complicated relations between concepts in knowledge representation. The associations among concepts can be classified into nine categories, such as inheritance, extension, tailoring, substitute, composition, decomposition, aggregation, specification, and instantiation as shown in Figure 6 and Table 2 (Wang, 2006e). In Figure 6, \( R = \{R^c, R^i, R^o\} \), and all nine associations describe composing rules among concepts, except instantiation that is a relation between a concept and a specific object.

**Definition 12.** A generic knowledge \( K \) is an \( n \)-nary relation \( R_k \) among a set of \( n \) multiple concepts in \( C \), that is:

\[
K = R_k : (\bigcup_{i=1}^{n} C_i) \rightarrow C
\]

where \( \bigcup_{i=1}^{n} C_i = C \), and \( R_k \in \mathcal{R} = \{\Rightarrow, \Rightarrow, \Rightarrow, \Rightarrow, \Rightarrow, \Rightarrow, \Rightarrow, \Rightarrow, \Rightarrow\} \).

In Definition 12, the relation \( R_k \) is one of the concept operations in CA as defined in Table 2 (Wang, 2006e) that serves as the knowledge composing rules.

**Definition 13.** A concept network \( CN \) is a hierarchical network of concepts interlinked by the set of nine associations \( \mathcal{R} \) defined in CA, that is:

---

**Figure 6.** The nine concept association operations as knowledge composing rules

![Diagram showing the nine concept association operations as knowledge composing rules.](image)
The Theoretical Framework of Cognitive Informatics

\[ CN = R \sum_{i,j}^{n} C_i \rightarrow \sum_{i,j}^{n} C_j \]  

(9)

where \( R \in \mathbb{R} \).

Because the relations between concepts are transitive, the generic topology of knowledge is a hierarchical concept network. The advantages of the hierarchical knowledge architecture \( K \) in the form of concept networks are as follows:

(a) **Dynamic**: The knowledge networks may be updated dynamically along with information acquisition and learning without destructing the existing concept nodes and relational links.

(b) **Evolvable**: The knowledge networks may grow adaptively without changing the overall and existing structure of the hierarchical network.

A summary of the algebraic relations and operations of concepts defined in CA are provided in Table 2.

**Real-Time Process Algebra (RTPA)**

A key metaphor in system modeling, specification, and description is that a software system can be perceived and described as the composition of a set of interacting processes. Hoare (1985), Milner (1989), and others developed various algebraic approaches to represent communicating and concurrent systems, known as process algebra. A **process algebra** is a set of formal notations and rules for describing algebraic relations of software processes. **Real-Time Process Algebra** (Wang, 2002b, 2005b) extends process algebra to time/event, architecture, and system dispatching manipulations in order to formally describe and specify architectures and behaviors of software systems. A process in RTPA is a computational operation that transforms a system from a state to another by changing its inputs, outputs, and/or internal variables. A process can be a single metaprocess or a complex process formed by using the process combination rules of RTPA known as process relations.

**Definition 14. Real-Time Process Algebra** is a set of formal notations and rules for describing algebraic and real-time relations of software processes.

RTPA models 17 metaprocesses and 17 process relations. A metaprocess is an elementary and primary process that serves as a common and basic building block for a software system. Complex processes can be derived from metaprocesses by a set of process relations that serves as process combinatory rules. Detailed semantics of RTPA may be referred to in Wang (2002b).

Program modeling is on coordination of computational behaviors with given data objects. Behavioral or instructive knowledge can be modeled by RTPA. A **generic program model** can be described by a formal treatment of statements, processes, and complex processes from the bottom-up in the program hierarchy.

**Definition 15.** A process \( P \) is a composed listing and a logical combination of \( n \) metatstatements \( p_i \) and \( p_j \), \( 1 \leq i < n, 1 < j \leq m = n+1 \), according to certain composing relations \( r_{ij} \), that is:

\[
P = \bigcap_{i=1}^{n} (p_i \rightarrow r_{ij} p_j), j = i + 1
\]

\[
= (\ldots (((p_i) \rightarrow r_{11} p_1) \rightarrow r_{23} p_3) \ldots r_{n-1,n} p_n)
\]

(10)

where the big-R notation (Wang, 2002b, 2006i) is adopted to describes the nature of processes as the building blocks of programs.

**Definition 16.** A program \( \Psi \) is a composition of a finite set of \( m \) processes according to the time-, event-, and interrupt-based process dispatching rules, that is:

\[
\Psi = \bigcap_{k=1}^{m} (@ e_k \rightarrow P_k)
\]

(11)

Equations 9.1 and 10.1 indicate that a program is an **embedded relational algebraic** entity. A
statement $p$ in a program is an instantiation of a metainstruction of a programming language that executes a basic unit of coherent function and leads to a predictable behavior.

**Theorem 5.** The embedded relational model (ERM) states that a software system or a program $\mathbb{P}$ is a set of complex embedded relational processes, in which all previous processes of a given process form the context of the current process, that is:

$$\mathbb{P} = \mathbb{R}[\langle e_i \rightharpoonup p_j \rangle]_{i=1}^{m}$$

(12)

ERM presented in Theorem 5 provides a unified mathematical model of programs (Wang, 2006a) for the first time, which reveals that a program is a finite and nonempty set of embedded binary relations between a current statement and all previous ones that formed the semantic context or environment of computing.

**Definition 17.** A metaprocess is the most basic and elementary processes in computing that cannot be broken up further. The set of metaprocesses $P$ encompasses 17 fundamental primitive operations in computing as follows:

$$P = \{:=, &, \Rightarrow, \Leftrightarrow, <, |>, |<, |<, \triangle, \downarrow, \!\!, \otimes, , \$\}$$

(13)

**Definition 18.** A process relation is a composing rule for constructing complex processes by using the metaprocesses. The process relations $R$ of RTPA are a set of 17 composing operations and rules to build larger architectural components and complex system behaviors using the metaprocesses, that is:

$$R = \{\rightarrow, \cap, |, \ldots, \emptyset, \cup, \Rightarrow, \|, \|\}$$

(14)

The definitions, syntaxes, and formal semantics of each of the metaprocesses and process relations may be referred to RTPA (Wang, 2002b, 2006f). A complex process and a program can be derived from the metaprocesses by the set of algebraic process relations. Therefore, a program is a set of embedded relational processes as described in Theorem 5.

A summary of the metaprocesses and their algebraic operations in RTPA are provided in Table 2.

**System Algebra (SA)**

Systems are the most complicated entities and phenomena in the physical, information, and social worlds across all science and engineering disciplines (Klir, 1992; von Bertalanffy, 1952; Wang, 2006d). Systems are needed because the physical and/or cognitive power of an individual component or person is not enough to carry out a work or solving a problem. An abstract system is a collection of coherent and interactive entities that has stable functions and clear boundary with external environment. An abstract system forms the generic model of various real world systems and represents the most common characteristics and properties of them.

**Definition 19.** System algebra is a new abstract mathematical structure that provides an algebraic treatment of abstract systems as well as their relations and operational rules for forming complex systems (Wang, 2006d).

Abstract systems can be classified into two categories known as the closed and open systems. Most practical and useful systems in nature are open systems in which there are interactions between the system and its environment. However, for understanding easily, the closed system is introduced first.
Definition 20. A closed system is a 4-tuple, that is:
\[ S = (C, R, B, \Omega) \] (15)

where
- \( C \) is a nonempty set of components of the system, \( C = \{c_1, c_2, \ldots, c_n\} \).
- \( R \) is a nonempty set of relations between pairs of the components in the system, \( R = \{r_1, r_2, \ldots, r_m\}, R \subseteq C \times C \).
- \( B \) is a set of behaviors (or functions), \( B = \{b_1, b_2, \ldots, b_p\} \).
- \( \Omega \) is a set of constraints on the memberships of components, the conditions of relations, and the scopes of behaviors, \( \Omega = \{\omega_1, \omega_2, \ldots, \omega_q\} \).

Most practical systems in the real world are not closed. That is, they need to interact with the external world known as the environment \( \Theta \) in order to exchange energy, matter, and/or information. Such systems are called open systems. Typical interactions between an open system and the environment are inputs and outputs.

Definition 21. An open system \( S \) is a 7-tuple, that is:
\[ S = (C, R, B, \Omega, \Theta) \]
\[ = (C, R^c, R^i, R^o, B, \Omega, \Theta) \] (16)

where the extensions of entities beyond the closed system are as follows:
- \( \Theta \) is the environment of \( S \) with a nonempty set of components \( C_\Theta \) outside \( C \).
- \( R^c \subseteq C \times C \) is a set of internal relations.
- \( R^i \subseteq C_\Theta \times C \) is a set of external input relations.
- \( R^o \subseteq C \times C_\Theta \) is a set of external output relations.

An open system \( S=(C, R^c, R^i, R^o, B, \Omega, \Theta) \) can be illustrated in Figure 7 (Wang, 2006d).

Theorem 6. The equivalence between open and closed systems states that an open system \( S \) is equivalent to a closed system \( \hat{S} \), or vice versa, when its environment or is conjoined, respectively, that is:
\[ \left\{ \begin{array}{l} \hat{S} = S \sqcup \Theta_S \\ S = \hat{S} \sqcup \Theta_{\hat{S}} \end{array} \right. \] (17)
According to Theorem 6, any subsystem of a closed system $\hat{S}$ is an open system $S$. That is, any supersystem $S$ of a given set of $n$ open systems $S_k$, plus their environments $\Theta_k$, $1 \leq k \leq n$, is a closed system. The algebraic relations and operations of systems in SA are summarized in Table 2.

**Theorem 7.** The Wang’s first law of system science, system fusion, states that system conjunction or composition between two systems $S_1$ and $S_2$ creates new relations $\Delta R_{12}$ and/or new behaviors (functions) $\Delta B_{12}$ that are solely a property of the new supersystem $S$ determined by the sizes of the two intersected component sets $\#(C_1)$ and $\#(C_2)$, that is:

$$
\Delta R_{12} = \#(R) - (\#(R_1) + \#(R_2)) = (\#(C_1 + C_2))^2 - (\#(C_1))^2 - (\#(C_2))^2 = 2 \#(C_1) \cdot \#(C_2)
$$

The discovery in Theorem 7 reveals that the mathematical explanation of system utilities is the newly gained relations $\Delta R_{12}$ and/or behaviors (functions) $\Delta B_{12}$ during the conjunction of two systems or subsystems. The empirical awareness of this key system property has been intuitively or qualitatively observed for centuries. However, Theorem 7 is the first rigorous explanation of the mechanism of system gains during system conjunctions and compositions. According to Theorem 7, the maximum incremental or system gain equals to the number of by-directly interconnection between all components in both $S_1$ and $S_2$, that is, $2 \#(C_1) \cdot \#(C_2)$.

**Theorem 8.** The Wang’s 2nd law of system science, the maximum system gain, states that work done by a system is always larger than any of its components, but is less than or is equal to the sum of those of its components, that is:

$$
W(S) \leq \sum_{i=1}^{n} W(e_i), \quad \eta \leq 1
$$

$$
W(S) > \max(W(e_i)), \quad e_i \in E_S
$$

There was a myth on an ideal system in conventional systems theory that supposes the work done by the ideal system $W(S)$ may be greater than the sum of all its components $W(e_i)$, that is:

$$
W(S) \geq \sum_{i=1}^{n} W(e_i)
$$

According to Theorems 7 and 8, the ideal system utility is impossible to achieve.

A summary of the algebraic operations and their notations in CA, RTPA, and SA is provided in Table 2. Details may be referred to in Wang (2006d, g).

**APPLICATIONS OF CI**

The last two sections have reviewed the latest development of fundamental researches in CI, particularly its theoretical framework and descriptive mathematics. A wide range of applications of CI has been identified in multidisciplinary and transdisciplinary areas, such as: (1) The architecture of future generation computers; (2) Estimation the capacity of human memory; (3) Autonomic computing; (4) Cognitive properties of information, data, knowledge, and skills in knowledge engineering; (5) Simulation of human cognitive behaviors using descriptive mathematics; (6) Agent systems; (7) CI foundations of software engineering; (8) Deductive semantics of software; and (9) Cognitive complexity of software.

**The Architecture of Future Generation Computers**

Conventional machines are invented to extend human physical capability, while modern information processing machines, such as computers, communication networks, and robots, are developed for extending human intelligence, memory, and the capacity for information processing (Wang, 2004). Recent advances in CI provide formal description of an entire set of cognitive processes of the brain (Wang et al.,
Table 2. Taxonomy of contemporary mathematics for knowledge representation and manipulation

<table>
<thead>
<tr>
<th>Operations</th>
<th>Concept Algebra</th>
<th>System Algebra</th>
<th>Real-Time Process Algebra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Super/sub relation</td>
<td>≻ / ≺</td>
<td>⊂ / ⊃</td>
<td>Assignment ≈ Sequence</td>
</tr>
<tr>
<td>Related / independent</td>
<td>←→ / ↔ / ↔</td>
<td>Evaluation</td>
<td>Jump ▼</td>
</tr>
<tr>
<td>Equivalent</td>
<td>= =</td>
<td>Addressing</td>
<td>Branch</td>
</tr>
</tbody>
</table>
| Consistent          | ≡               | Memory allocation ≡ Switch | ...
| Overlapped          | Π               | Memory release ≡ While-loop |
| Conjunction         | +               | Read            | Repeat-loop               |
| Elicitation         | *               | Write           | For-loop                  |
| Comparison          | ~               | Input           | Recursion                 |
| Definition          | △               | Output          | Procedure call            |
| Difference          | □               | Timing          | Parallel                  |
| Inheritance         | ⇒ ⇒             | Duration        | Concurrence               |
| Extension           |                 | Increase        | Interleave                |
| Tailoring           |                 | Decrease        | Pipeline                  |
| Substitute           |                 | Exception       | Interrupt                 |
| Composition         | ⊔ ⊔             | Skip            | Time-driven dispatch      |
| Decomposition       | ⊔ ⊔             | Stop            | Event-driven dispatch     |
| Aggregation/        | ≜               | System          | Interrupt-driven dispatch |
| generalization      |                 |                 |                           |
| Specification       | ⊥ ⊥             |                 |                           |
| Instantiation       | ⊞ ⊞             |                 |                           |

2006). The fundamental research in CI also creates an enriched set of contemporary denotational mathematics (Wang, 2006c), for dealing with the extremely complicated objects and problems in natural intelligence, neural informatics, and knowledge manipulation.

The theory and philosophy behind the next generation computers and computing methodologies are CI (Wang, 2003b, 2004). It is commonly believed that the future-generation computers, known as the cognitive computers, will adopt non-von Neumann (von Neumann, 1946) architectures. The key requirements for implementing a conventional stored-program controlled computer are the generalization of common computing architectures and the computer is able to interpret the data loaded in memory as computing instructions. These are the essences of stored-program...
controlled computers known as the von Neumann (1946) architecture. Von Neumann elicited five fundamental and essential components to implement general-purpose programmable digital computers in order to embody the concept of stored-program-controlled computers.

**Definition 22.** A von Neumann Architecture (VNA) of computers is a 5-tuple that consists of the components: (a) the arithmetic-logic unit (ALU), (b) the control unit (CU) with a program counter (PC), (c) a memory (M), (d) a set of input/output (I/O) devices, and (e) a bus (B) that provides the data path between these components, that is:

$$VNA \triangleq (ALU, CU, M, I/O, B)$$

**Definition 23.** Conventional computers with VNA are aimed at stored-program-controlled data processing based on mathematical logic and Boolean algebra.

A VNA computer is centric by the bus and characterized by the all purpose memory for both data and instructions. A VNA machine is an extended Turing machine (TM), where the power and functionality of all components of TM including the control unit (with wired instructions), the tape (memory), and the head of I/O, are greatly enhanced and extended with more powerful instructions and I/O capacity.

**Definition 24.** A Wang Architecture (WA) of computers, known as the Cognitive Machine as shown in Figure 8, is a parallel structure encompassing an Inference Engine (IE) and a Perception Engine (PE) (Wang, 2006b, g), that is:

$$WA \triangleq (IE \parallel PE)$$

$$= ( KMU// The knowledge manipulation unit$$

$$\parallel BMU// The behavior manipulation unit$$

$$\parallel EMU // The experience manipulation unit$$

$$\parallel SMU// The skill manipulation unit$$

$$\parallel ( BPU // The behavior perception unit$$

$$\parallel EPU // The experience perception unit$$

$$)$$

$$CM = IE \parallel PE$$

$$= ( KMU// The knowledge manipulation unit$$

$$\parallel BMU// The behavior manipulation unit$$

$$\parallel EMU // The experience manipulation unit$$

$$\parallel SMU// The skill manipulation unit$$

$$\parallel ( BPU // The behavior perception unit$$

$$\parallel EPU // The experience perception unit$$

$$)$$

(21)
As shown in Figure 8 and Equation 21, WA computers are not centered by a CPU for data manipulation as the VNA computers do. The WA computers are centered by the concurrent IE and PE for cognitive learning and autonomic perception based on abstract concept inferences and empirical stimuli perception. The IE is designed for concept/knowledge manipulation according to concept algebra (Wang, 2006e), particularly the nine concept operations for knowledge acquisition, creation, and manipulation. The PE is designed for feeling and perception processing according to RTPA (Wang, 2002b) and the formally described cognitive process models of the perception layers as defined in the LRMB model (Wang et al., 2006).

**Definition 25.** Cognitive computers with WA are aimed at cognitive and perceptive concept/knowledge processing based on contemporary denotational mathematics, that is, CA, RTPA, and SA.

As that of mathematical logic and Boolean algebra are the mathematical foundations of VNA computers. The mathematical foundations of WA computers are based on denotational mathematics (Wang, 2006b, c). As described in the LRMB reference model (Wang et al., 2006), since all the 37 fundamental cognitive processes of human brains can be formally described in CA and RTPA (Wang, 2002b, 2006e). In other words, they are simulatable and executable by the WA-based cognitive computers.

**Estimation of the Capacity of Human Memory**

Despite the fact that the number of neurons in the brain has been identified in cognitive and neural sciences, the magnitude of human memory capacity is still unknown. According to the OAR model, a recent discovery in CI is that the upper bound of memory capacity of the human brain is in the order of $10^{8.432}$ bits (Wang et al., 2003). The determination of the magnitude of human memory capacity is not only theoretically significant in CI, but also practically useful to unveil the human potential, as well as the gaps between the natural and machine intelligence. This result indicates that the next generation computer memory systems may be built according to the OAR model rather than the traditional container metaphor, because the former is more powerful, flexible, and efficient to generate a tremendous memory capacity by using limited number of neurons in the brain or hardware cells in the next generation computers.

**Autonomic Computing**

The approaches to implement intelligent systems can be classified into those of biological organisms, silicon automata, and computing systems. Based on CI studies, *autonomic computing* (Wang, 2004) is proposed as a new and advanced computing technique built upon the routine, algorithmic, and adaptive systems as shown in Table 3.

The approaches to computing can be classified into two categories known as imperative and autonomic computing. Corresponding to these,

---

**Table 3. Classification of computing systems**

<table>
<thead>
<tr>
<th>Event (I)</th>
<th>Behavior (O)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>Routine</td>
</tr>
<tr>
<td>Variable</td>
<td>Algorithmic</td>
</tr>
<tr>
<td>Type of behavior</td>
<td>Deterministic</td>
</tr>
<tr>
<td></td>
<td>Nondeterministic</td>
</tr>
</tbody>
</table>
computing systems may be implemented as imperative or autonomic computing systems.

**Definition 26.** An imperative computing system is a passive system that implements deterministic, context-free, and stored-program controlled behaviors.

**Definition 27.** An autonomic computing system is an intelligent system that autonomously carries out robotic and interactive actions based on goal- and event-driven mechanisms.

The imperative computing system is a traditional passive system that implements deterministic, context-free, and stored-program controlled behaviors, where a behavior is defined as a set of observable actions of a given computing system. The autonomic computing system is an active system that implements nondeterministic, context-dependent, and adaptive behaviors, which do not rely on instructive and procedural information, but are dependent on internal status and willingness that formed by long-term historical events and current rational or emotional goals.

The first three categories of computing techniques as shown in Table 3 are imperative. In contrast, the autonomic computing systems are an active system that implements nondeterministic, context-sensitive, and adaptive behaviors. Autonomic computing does not rely on imperative and procedural instructions, but are dependent on perceptions and inferences based on internal goals as revealed in CI.

### Cognitive Properties of Knowledge

Almost all modern disciplines of science and engineering deal with information and knowledge. According to CI theories, cognitive information may be classified into four categories known as knowledge, behaviors, experience, and skills as shown in Table 4.

**Definition 28.** The taxonomy of cognitive information is determined by its types of inputs and outputs to and from the brain during learning and information processing, where both inputs and outputs can be either abstract information (concept) or empirical information (actions).

It is noteworthy that the approaches to acquire knowledge/behaviors and experience/skills are fundamentally different. The former may be obtained either directly based on hands-on activities or indirectly by reading, while the latter can never be acquired indirectly.

According to Table 4, the following important conclusions on information manipulation and learning for both human and machine systems can be derived.

**Theorem 9.** The principle of information acquisition states that there are four sufficient categories of learning known as those of knowledge, behaviors, experience, and skills.

Theorem 9 indicates that learning theories and their implementation in autonomic and intelligent systems should study all four categories of

### Table 4. Types of cognitive information

<table>
<thead>
<tr>
<th>Type of Input</th>
<th>Type of Output</th>
<th>Ways of Acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract Concept</td>
<td>Knowledge</td>
<td>Direct or indirect</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Empirical Action</td>
<td>Experience</td>
</tr>
</tbody>
</table>

51
cognitive information acquisitions, particularly behaviors, experience, and skills rather than only focusing on knowledge.

**Corollary 3.** All the four categories of information can be acquired directly by an individual.

**Corollary 4.** Knowledge and behaviors can be learnt indirectly by inputting abstract information, while experience and skills must be learned directly by hands-on or empirical actions.

The above theory of CI lays an important foundation for learning theories and pedagogy (Wang, 2004, 2006e). Based on the fundamental work, the IE and PE of cognitive computers working as a virtual brain can be implemented on WA-based cognitive computers and be simulated on VNA-based conventional computers.

**Simulation of Human Cognitive Behaviors using the Contemporary Mathematics**

The contemporary denotational mathematics as described in The Denotational Mathematics for CI section, particularly CA and RTPA, may be used to simulate the cognitive processes of the brain as modeled in LRMB (Wang et al., 2006). Most of the 37 cognitive processes identified in LRMB, such as the learning (Wang, 2006c), reasoning (Wang, 2006b), decision making (Wang et al., 2004), and comprehension (Wang & Gafurov, 2003) processes, have been rigorously modeled and described in RTPA and CA. Based on the fundamental work, the inference engineering and perception engine of a virtual brain can be implemented on cognitive computers or be simulated on conventional computers. In the former case, a working prototype of a fully autonomic computer will be realized on the basis of CI theories.

**Agent Systems**

**Definition 29.** A *software agent* is an intelligent software system that autonomously carries out robotic and interactive applications based on goal-driven mechanisms (Wang, 2003c).

Because a software agent may be perceived as an application-specific virtual brain (see Theorem 3), behaviors of an agent are mirrored human behaviors. The fundamental characteristics of agent-based systems are autonomic computing, goal-driven action-generation, knowledge-based machine learning. In recent CI research, perceptivity is recognized as the sixth sense that serves the brain as the thinking engine and the kernel of the natural intelligence. Perceptivity implements self-consciousness inside the abstract memories of the brain. Almost all cognitive life functions rely on perceptivity such as consciousness, memory searching, motivation, willingness, goal setting, emotion, sense of spatiality, and sense of motion. The brain may be stimulated by external and internal information, which can be classified as willingness-driven (internal events such as goals, motivation, and emotions), event-driven (external events), and time-driven (mainly external events triggered by an external clock). Unlike a computer, the brain works in two approaches: the internal willingness-driven processes, and the external event- and time-driven processes. The external information and events are the major sources that drive the brain, particularly for conscious life functions.

Recent research in CI reveals that the foundations of agent technologies and autonomic computing are CI, particularly goal-driven action generation techniques (Wang, 2003c). The LRMB model (Wang et al., 2006) described in the Layered Reference Model of the Brain section may be used as a reference model for agent-based technologies. This is a fundamental view toward
The formal description and modeling of architectures and behaviors of agent systems, which are created to do something repeatable in context, to extend human capability, reachability, and/or memory capacity. It is found that both human and software behaviors can be described by a 3-dimensional representative model comprising action, time, and space. For agent system behaviors, the three dimensions are known as mathematical operations, event/process timing, and memory manipulation (Wang, 2006g). The 3-D behavioral space of agents can be formally described by RTPA that serves as an expressive mathematical means for describing thoughts and notions of dynamic system behaviors as a series of actions and cognitive processes.

**CI Foundations of Software Engineering**

Software is an intellectual artifact and a kind of instructive information that provides a solution for a repeatable computer application, which enables existing tasks to be done easier, faster, and smarter, or which provides innovative applications for the industries and daily life. Large-scale software systems are highly complicated systems that have never been handled or experienced precedent by mankind.

The fundamental cognitive characteristics of software engineering have been identified as follows (Wang, 2006g):

- The inherent complexity and diversity
- The difficulty of establishing and stabilizing requirements
- The changeability or malleability of system behavior
- The abstraction and intangibility of software products
- The requirement of varying problem domain knowledge
- The non-deterministic and polysolvability in design
- The polyglotics and polymorphism in implementation
- The dependability of interactions among software, hardware, and human beings

The above list forms a set of fundamental constraints for software engineering, identified as the cognitive constraints of intangibility, complexity, indeterminacy, diversity, polymorphism, inexpressiveness, inexplicit embodiment, and unquantifiable quality measures (Wang, 2006g).

A set of psychological requirements for software engineers has been identified, such as:

(a) Abstract-level thinking; (b) Imagination of dynamic behaviors with static descriptions; (c) Organization capability; (d) Cooperative attitude in team work; (e) Long-period focus of attentions; (f) Preciseness; (g) Reliability; and (h) Expressive capability in communication.

**Deductive Semantics of Software**

Deduction is a reasoning process that discovers new knowledge or derives a specific conclusion based on generic premises such as abstract rules or principles. In order to provide an algebraic treatment of the semantics of program and human cognitive processes, a new type of formal semantics known as deductive semantics is developed (Wang, 2006f, g).

**Definition 30.** Deductive semantics is a formal semantics that deduces the semantics of a program from a generic abstract semantic function to the concrete semantics, which are embodied onto the changes of status of a finite set of variables constituting the semantic environment of computing (Wang, 2006g).

**Theorem 10.** The semantics of a statement \( p, \Theta(p) \) on a given semantic environment \( \Theta \) in deductive semantics is a double partial differential of the semantic function, \( f_{\Theta(p)}(p) = f_{\Theta}(p) \), on the sets of variables \( S \) and executing steps \( T \), that is:
Theoretical Framework of Cognitive Informatics

\[ \theta (p) = \frac{\partial^2}{\partial t \partial s} f_s (p) = \frac{\partial^2}{\partial t \partial s} v_s (t, s) \]

\[ = R \sum_{i=0}^{\#T(p)} R \sum_{j=1}^{\#S(p)} v_p (t_i, s_j) \]

\[ = R \sum_{i=0}^{\#T(p)} R \sum_{j=1}^{\#S(p)} v_p (t_i, s_j) \]

\[ = \left( \begin{array}{cccc} t_0 & s_1 & s_2 & \cdots & s_m \\ v_{01} & v_{02} & \cdots & v_{0m} \\ t_1 & s_1 & s_2 & \cdots & s_m \\ v_{11} & v_{12} & \cdots & v_{1m} \end{array} \right) \]  

(22)

where \( t \) denotes the discrete time immediately before and after the execution of \( p \) during \( (t_0, t_1] \), and \( \# \) is the cardinal calculus that counts the number of elements in a given set, that is \( n = \#T(p) \) and \( m = \#S(p) \).

The first partial differential in Equation 22 selects all related variable \( S(p) \) of the statement \( p \) from \( \Theta \). The second partial differential selects a set of discrete steps of \( p \)’s execution \( T(p) \) from \( \Theta \). According to Theorem 10, the semantics of a statement can be reduced onto a semantic function that results in a 2-D matrix with the changes of values for all variables over time along program execution.

Deductive semantics perceives that the carriers of software semantics are a finite set of variables declared in a given program. Therefore, software semantics can be reduced onto the changes of values of these variables. The deductive mathematical models of semantics and the semantic environment at various composing levels of systems are formally described. Properties of software semantics and relationships between the software behavioral space and the semantic environment are discussed. Deductive semantics is applied in the formal definitions and explanations of the semantic rules of a comprehensive set of software static and dynamic behaviors as modeled in RTPA. Deductive semantics can be used to define abstract and concrete semantics of software and cognitive systems, and facilitate software comprehension and recognition by semantic analyses.

Cognitive Complexity of Software

The estimation and measurement of functional complexity of software are an age-long problem in software engineering. The cognitive complexity of software (Wang, 2006) is a new measurement for cross-platform analysis of complexities, sizes, and comprehension effort of software specifications and implementations in the phases of design, implementation, and maintenance in software engineering. This work reveals that the cognitive complexity of software is a product of its architectural and operational complexities on the basis of deductive semantics and the abstract system theory. Ten fundamental basic control structures (BCSs) are elicited from software architectural/behavioral specifications and descriptions. The cognitive weights of those BCSs are derived and calibrated via a series of psychological experiments. Based on this work, the cognitive complexity of software systems can be rigorously and accurately measured and analyzed. Comparative case studies demonstrate that the cognitive complexity is highly distinguishable in software functional complexity and size measurement in software engineering.

On the basis of the ERM model described in Theorem 5 and the deductive semantics of software presented in The deductive semantics of software section, the finding on the cognitive complexity of software is obtained as follows.

**Theorem 11.** The sum of the cognitive weights of all \( r_{ij} \); \( w(r_{ij}) \), in the ERM model determines the operational complexity of a software system \( C_{op} \), that is:

\[ C_{op} = \sum_{i=1}^{n} w(r_{ij}), j = i + 1 \]  

(23)

A set of psychological experiments has been carried out in undergraduate and graduate classes in software engineering. Based on 126 experiment results, the equivalent cognitive weights of the
The Theoretical Framework of Cognitive Informatics

10 fundamental BCSs are statistically calibrated as summarized in Table 5 (Wang, 2006j), where the relative cognitive weight of the sequential structures is assumed one, that is, \( w_i = 1 \).

According to deductive semantics, the complexity of a software system, or its semantic space, is determined not only by the number of operations, but also by the number of data objects.

**Theorem 12.** The cognitive complexity \( C_c(S) \) of a software system \( S \) is a product of the operational complexity \( C_{op}(S) \) and the architectural complexity \( C_a(S) \), that is:

\[
C_c(S) = C_{op}(S) \cdot C_a(S)
\]

Based on Theorem 11, the cognitive complexities of four typical software components (Wang, 2006j) have been comparatively analyzes as summarized in Table 6. For enabling comparative analyses, data based on existing complexity measures, such as time, cyclomatic, and symbolic (LOC) complexities, are also contrasted in Table 6.

Observing Table 6 it can be seen that the first three traditional measurements cannot actually reflect the real complexity of software systems in software design, representation, cognition, comprehension, and maintenance. It is found that (a) Although four example systems are with similar symbolic complexities, their operational and functional complexities are greatly different. This indicates that the symbolic complexity cannot be used to represent the operational or functional complexity of software systems. (b) The symbolic complexity (LOC) does not represent the throughput or the input size of problems. (c) The time complexity does not work well for a system where there are no loops and dominate operations, because in theory that all statements

<table>
<thead>
<tr>
<th>BCS</th>
<th>RTPA Notation</th>
<th>Description</th>
<th>Calibrated cognitive weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \rightarrow )</td>
<td>Sequence</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>(</td>
<td>)</td>
<td>Branch</td>
</tr>
<tr>
<td>3</td>
<td>( \ldots</td>
<td>\ldots )</td>
<td>Switch</td>
</tr>
<tr>
<td>4</td>
<td>( R^i )</td>
<td>For-loop</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>( R^* )</td>
<td>Repeat-loop</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>( R^* )</td>
<td>While-loop</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>( \rightarrow )</td>
<td>Function call</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>( \circ )</td>
<td>Recursion</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>( \parallel ) or ( \square )</td>
<td>Parallel</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>( \odot )</td>
<td>Interrupt</td>
<td>22</td>
</tr>
</tbody>
</table>
in linear structures are treated as zero in this measure no matter how long they are. In addition, time complexity cannot distinguish the real complexities of systems with the same asymptotic function, such as in Case 2 (IBS (b)) and Case 3 (Maxfinder). (d) The cognitive complexity is an ideal measure of software functional complexities and sizes, because it represents the real semantic complexity by integrating both the operational and architectural complexities in a coherent measure. For example, the difference between IBS(a) and IBS(b) can be successfully captured by the cognitive complexity. However, the symbolic and cyclomatic complexities cannot identify the functional differences very well.

**CONCLUSIONS**

This article has presented an intensive survey of the recent advances and groundbreaking studies in Cognitive informatics, particularly its theoretical framework, denotational mathematics, and main application areas. CI has been described as a new discipline that studies the natural intelligence and internal information processing mechanisms of the brain, as well as processes involved in perception and cognition. CI is a new frontier across disciplines of computing, software engineering, cognitive sciences, neuropsychology, brain sciences, and philosophy in recent years. It has been recognized that many fundamental issues in knowledge and software engineering are based on the deeper understanding of the mechanisms of human information processing and cognitive processes.

A coherent set of theories for CI has been described in this article, such as the Information-Matter-Energy model, Layered Reference Model of the Brain, the OAR model of information representation, Natural Intelligence vs. Artificial Intelligence, Autonomic Computing vs. imperative computing, CI laws of software, mechanisms of human perception processes, the cognitive processes of formal inferences, and the formal knowledge system. Three contemporary mathematical means have been created in CI known as the denotational mathematics. Within the new forms of denotational mathematical means for CI, Concept Algebra has been designed to deal with the new abstract mathematical structure of concepts and their representation and manipulation in learning and knowledge engineering. Real-Time Process Algebra has been developed as an expressive, easy-to-comprehend, and language-independent notation system, and a specification and refinement method for software system behaviors description and specification. System Algebra has been created to the rigorous treatment of abstract systems and their algebraic relations and operations.

A wide range of applications of CI has been identified in multidisciplinary and transdisciplinary areas, such as the architecture of future generation computers, estimation the capacity of human memory, autonomic computing, cognitive properties of information, data, knowledge,
The Theoretical Framework of Cognitive Informatics

and skills in knowledge engineering, simulation of human cognitive behaviors using descriptive mathematics, agent systems, CI foundations of software engineering, deductive semantics of software, and cognitive complexity of software systems.

ACKNOWLEDGMENT

The author would like to acknowledge the Natural Science and Engineering Council of Canada (NSERC) for its support to this work. The author would like to thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES


The Theoretical Framework of Cognitive Informatics


actions on Systems, Man, and Cybernetics (Part C), 36(2), 121-123.


*This work was previously published in International Journal of Cognitive Informatics and Natural Intelligence, Vol. 1, Issue 1, edited by Y. Wang, pp. 1-27, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 1.5
Human Factors in Public Information Systems

Karl W. Sandberg
Mid Sweden University, Sweden

Pan Yan
Luleå University of Technology, Sweden

INTRODUCTION

News reports do not frequently mention many problems or accidents caused by human error. The specialty of human factors seeks to avoid human error by making certain that computers and all other equipment are designed to be easy to understand and use; costly human errors are thus minimised. This article provides a basic overview of the subject of human factors as it pertains to problem and error avoidance in computerised public information systems. When computer/system design does not adequately consider human capability, the performance of the computer/system and the user will be below desired levels.

BACKGROUND: HUMAN FACTORS, THE DISCIPLINE

Human factors has helped to make information systems less costly and easier to use. Some background on this discipline will help to show the relationship it has with information systems.

Human factors is the science of the relationships between humans, their activities, and the context in which those activities take place. The specialty of human factors came into being during WWII when it was found that the sophisticated equipment being designed did not perform as expected; there was a strong and immediate need to understand why many humans were unable to effectively use equipment and/or systems. Teams of researchers learned that the design strategies...
used did not consider human performance limitations—for example, some designs presented too much information at the same time or in the wrong order for humans to be able to successfully operate controls. Or, the arrangement of controls made them difficult to reach quickly and easily. From this discovery came the concept that human users, their work activities, and the contexts of the activities had to be thought of as different parts of a whole system and that each depends upon the other for successful operation (Bailey, 1996).

After WWII the discipline of human factors became a specialised knowledge area as it became apparent that the human element of any system had to be considered if the capabilities of new technologies were to be efficiently exploited. The older strategy of modifying designs over a long period of time through user experiences was inadequate; rates of change had become so rapid that products were obsolete before improvements could be added. Now, the strategy often used by successful design environments is to include human factors in design and development. When properly managed, products or systems that use human factors knowledge are more efficient, safer, and more pleasing to use because they are designed to accommodate human performance capabilities (Norman, 1988).

Human factors is an extremely broad technical and scientific discipline; founders of the first national and international human factors organisations came from such diverse fields as engineering, design, education, computer technology, psychology, and medicine. Through its diversity human factors is able to draw upon and combine knowledge from any area when working with human and system performance issues. Due to the complexity of human behaviour, human factors specialists emphasise in their work an iterative empirical approach. First, an initial recommendation or interface design is made and then laboratory or field studies are conducted to test this initial design (the prototype). Changes are made when deficits are identified; modifications are made; and further testing is then performed. This process continues until significant problems are no longer found. Finally, validation is achieved through observation in the field after system deployment.

This emphasis on empirical work tends to shape how human factors specialists perform their roles. Irrespective of the specific methodology chosen for gathering data about tasks, users, and the use of products, human factors work tends to result in product improvements likely to be economical, easy, and efficient to use from the beginning of use; the cost of and need to go back and fix problems when human factors is not used is avoided.

Human factors can also be called ergonomics. As the term “human factors” is in more common usage in the computer field, it is used for this article.

Human Factors and Computers

As with other technologies, WWII helped to stimulate the development of computers. The first models of computer used to manage information were primitive in comparison to the computers of today. They were designed by scientists and engineers to be used by scientists and engineers. These computer systems were difficult to operate and had to be closely watched during operation; users had to understand the technology for successful operation and tolerate problems, as technology-advanced computers became easier to operate. With today’s systems, the typical user does not need to understand technology; what is important is that a user understands how to tell a computer exactly what information is needed.

This progress in computer technology means that average users of computers and information technology (IT) are no longer just scientists or engineers; instead, they can be anyone who understands how to operate the computer information system they are using. While computer and IT systems are much easier to use, this does not
mean that there are not challenges faced. When working with information stored electronically or operating something with the help of a computer, problems can occur when a user misunderstands information presented to them, or they fail to correctly tell the computing system what is wanted. Literature on computers and electronic technology describes instances where there were serious negative consequences when information presented by a computer was misunderstood or systems were mistakenly told to do something different than what was wanted. The shooting down of Korean Airlines flight 007 in 1983 demonstrates some of the problems that can occur between the operator and a computer. Before take-off a crew member entered wrong data into an on-board computer. A warning was ignored as it appeared to be caused by another factor other than the entry of incorrect data. Subsequently the incorrect data caused the pilot to fly off course and into a sensitive area where the aircraft was shot down by a missile. Had the system been designed differently, so that the human activity context was better considered, then it would be likely that a tragic chain of events would not have been started.

To prevent such events from occurring, an emphasis on understanding how to design systems, where it is difficult for people to make mistakes or misunderstand, has emerged as a specialty within the field of human factors. The concept of human-computer interaction, or HCI, as a specialty within human factors emerged. In simple terms, HCI is the search for ways to make computers and complex information services accessible, usable, and acceptable to the non-specialist user.

At first, the emphasis of HCI was on physical aspects such as the size, location, and qualities of computer screens and controls. How people think when using computing and IT systems was added as it became more and more evident that how users think and understand is extremely critical in computer design. The subspecialties of cognitive engineering and usability followed—the former concerns understanding how people think when understanding systems, and the latter concerns how well people can understand and use systems.

Human factors and its related disciplines are now recognised as an important part of the computing system design process. Most computer and related system design approaches try to consider the needs and capabilities of users. With new products and technologies constantly entering the market, new problems are continually encountered—this underscores a continuing need to address human factors when designing computers and computing systems (Cooper, 1999).

The migration of paper-based forms to computer-based public information systems is the type of problem now faced. Work currently in progress in Sweden has found that the way a form looks on paper cannot be used on a computer screen. A computerised system had been developed that used exact copies of paper forms on computer screens. User error rates were extremely high, and the computer system was more costly to operate than its paper-based predecessor. The computer system had been designed without involvement of human factors specialists. Human factors specialists were asked to help correct system problems. The initial results of the human factors assessment reduced error rates and costs; work continues to bring performance to a desired level. The human factors specialists questioned and observed users. From those, it was possible to identify simple solutions to performance problems.

Public Information Systems, Computers, and Human Factors

Since public and private information systems use the same technology and function to exchange information between users and systems, is there any difference between the two? Should they be treated differently? Do the two have different needs?
Questions about difference and how human factors is connected are perhaps best answered by looking at the populations served and the reasons for the creation of any particular system. Private information systems generally serve segments of entire populations (exclusive) and make little or no attempt to accommodate those who lack needed skills. In contrast, public systems must at least have the goal of being usable by every member of a population (inclusive). The exclusive nature of private information systems means that during development and design there is a focus on meeting the needs of a target group that is part of an entire population. In contrast, public information systems generally must be developed and designed to at least have the goal of accessibility to an entire population. A second major difference between the two is that private information systems are oriented towards generating at least enough revenue to meet expenses, while public information systems are oriented towards providing service (which may or may not generate revenue).

While these differences may be subtle in some circumstances (e.g., a system used by a private, non-profit organisation or a system used by a fee-based public entity), it can be argued that, in some form, an inclusive or exclusive orientation is present.

How is this difference important with respect to human factors and public information systems? It is important because public information systems must be designed to be usable by the public—which includes everyone in a population. The goal of accessibility to an entire population means having to design for all levels of computer skills, education, and ability. A private information system is designed to be used by persons who are already familiar with computers and, if specialised knowledge/skills are needed, training is provided.

Any entity responsible for a public information system benefits from incorporating human factors knowledge and skills into any information system development or operation. The multiple levels of needs of users make understanding this diversity essential. Without human factors there is a much higher risk that the best design alternatives will not be found and selected—such as trying to move from paper forms to computer-based forms without considering the differences between entering data on paper and entering the same data onto a computer screen. As described in greater detail in the following section, human factors practices make certain that people can easily understand and use a system.

**HUMAN FACTORS AS A KEY ELEMENT IN PUBLIC SYSTEM DESIGN**

Including human factors in the earliest stages of any design process ensures that a product is defined in terms of user needs. Human factors input at these stages also helps project management to redirect resources to approaches that are most likely to have the greatest user benefit. Human factor specialist involvement throughout design processes results in catching problems early on; thus helping to reduce overall project costs. Finally, the ability to anticipate user responses continues to improve as successive prototypes more closely approximate the final product and actual field conditions. During design and development processes, human factors specialists base user interface and system design recommendations on an understanding of user needs, goals, characteristics, and expectations. Systems designers then work with human factors specialists to translate these stated needs into specific technical design requirements that may be implemented within cost targets. As noted earlier, the human factors specialist can work with public information system development and design processes to balance usability requirements with user constraints, system requirements, and development costs. The result can be that the final product functions as desired and in ways that are easily usable (Bailey, 1996).
PLANNING FOR HUMAN FACTORS DURING PUBLIC INFORMATION SYSTEM DEVELOPMENT

Information systems’ planning is a common problem for organisations, and planning for human factors is a particular problem within the overall problem of design (Kuark, 1996). The traditional approach to planning IT systems is called the systems development life cycle (SDLC) method (Davis & Olson, 1985). This approach can be summarised as a linear checklist of activities undertaken in a series of phases that begins with planning and ends with system termination. To put it another way, SDLC is a planning process that separates an IT system problem into its constituent parts, treats each step separately, and then unifies the parts into a whole at the end (Lange-fors, 1973). This approach has been criticised as too concentrated on technology at the expense of users and product performance (e.g., Noyes et al., 1996). To keep the valuable parts of the SDLC method there is a movement to integrate human factors into SDLC (Zhang, Carey, Te’eni, & Tremaine, 2005). Whether integrated in with SDLC or another methodology, HCI development and design is the process of creating quickly all or part of a system in the form of a simulation or prototype (e.g., on paper or a nonworking sample displayed on a computer screen). The advantage of this method is that it allows user feedback early enough to accommodate major structural changes before extensive development investments are committed. This use of prototyping is a team process which helps to overcome the limitations of SDLC. The early SDLC model has been described as too linear, lacking a group planning orientation and an actual reflection of real world conditions (e.g., Wanninger & Dickson, 1992). Now the emphasis is on using the systematic features of SDLC from a human factors perspective; technical decisions are based on an understanding of human needs and capabilities (e.g., Zhang et al., 2005). The following section on user-centred design describes the most commonly used model promoted by human factors professionals and user-oriented designers.

USER-CENTRED DESIGN

User-centred design (UCD) refers to the design of interaction between users and the system, called interaction design (Preece, Rogers, & Sharp, 2002). It models a system from a user perspective and focuses on the usability of an interactive software system. The core objective is to effectively support users in executing their tasks (Earthly, 2001).

Usability is recognised as one of the most important quality characteristics of software intensive systems and products. Usability gives many benefits including “increased productivity, enhanced quality of work, improved user satisfaction, reductions in support and training costs and improved user satisfaction” (ISO13407, 1999).

The prevailing paradigm of developing usable products and systems (or UCD) is that usable products are created through processes of user-centred design. The UCD process model is illustrated in Figure 1. Achieving quality in use requires this type of user-centred design process and the use of appropriate usability evaluation techniques. Usability is defined as a high level quality objective: to achieve effectiveness, efficiency, and satisfaction. This requires not only ease of use but also appropriate functionality, reliability, computer performance, satisfaction, comfort, and so on.

UCD and Usability

The role of UCD in a development effort is to deal with user needs in a way that highlights potential problems and ensures that products, systems, and services will succeed because they are based on appropriate UCD and usability. The applied discipline of cognitive engineering leans heavily on
the knowledge and methods developed by human experimental psychology and, to a lesser extent, on applied physiology and related human sciences to carry this goal out.

Usability Facilities

Usability laboratories have become an important tool. These laboratories, together with appropriate field-kit evaluations, enable us to meet the growing number of requests for usability support, chiefly user-based evaluations of development prototypes and competitors’ products. Together with the appropriate development processes, laboratory studies help predict and deal with specific usability problems during product development. Such a laboratory also provides an extremely powerful tool to identify and publicise general usability issues that affect overall product use. Rubin (1994) describes usability testing in detail and gives examples of typical usability laboratories.

IMPLEMENTATION ISSUES

Although improved technology will always have a critical role in system design, it is often the case that human factors considerations will take up the majority of the time of IT designers and managers. Much is known about the nature of human error, the conditions that encourage error, and hardware/software designs that are error resistant (Norman, 1990), and much of this is now well recognised in IT training literature (Salvendy, 1997). While human factors is recognised and used by many information system developers and designers, there are those who take the perspective that users who want access must accommodate themselves to an information system.

The attitude of the employees’ response for administering a public information system is also critical for success. New users or those with problems may need to ask an employee (a human-human interface) for help. Employees who have a negative attitude will too often give the user a negative attitude towards the system (Cialdini, 1993). Including employees in processes will help to avoid the problem of employees undermining the operation of a public information system; following are some points to use in the avoidance of employee resistance.

Cognitive Aspects of Implementation

Overcoming employee resistance is a problem in cognitive change. Human factors texts and refer-
ence books describe how computerisation alters the flow and content of information, changes job content, or affects relationships between organisational members (e.g., Bailey, 1996; Dix, Finlay, Abowd, & Beale, 1998; Helander, 1997). A key finding was that the primary issue was not computing itself, but rather perceptions of the method of implementing it. It is precisely in this area that human factors become important by including assessment of organisations during public information system changes.

**Cognitive Perception**

In IT projects there are four common cognitive perceptions that must be overcome by employees: (1) automation is mandated arbitrarily; (2) the new computer system will be unreliable; (3) the new system will increase rather than decrease the work burden; and (4) they (employees) will neither understand a system nor be able to operate it. Free and open communications throughout an IT implementation process is an important avenue for reducing employee resistance coming from these issues. Negative perceptions regarding the introduction of IT may be overcome to a significant extent through the encouragement of employee decision making during an IT introduction process.

**Attitudes**

Although user attitudes tend to be against change, once new technology is seen as bringing desired benefits, attitudes will begin to adapt. In particular, acceptance is much more likely when computing is presented as complementary to human skills, enhancing rather than replacing them (Petheram, 1989; Rosenbrock, 1981). Changing employee attitudes is all the more important since word of mouth is a critical aspect of the process by which technological innovations are spread (Czepiel, 1974). Dickson, a pioneer in the study of human relations and IT noted that, when it comes to management information systems, it is essential to involve people as a component, knowledge of their behaviour as a part of the system is important to the success of the entire field (Dickson, 1968).

The human factor in IT involves both teamwork and leadership. If a leader attempts to bring about IT change in an authoritarian manner, the result may be resistance, lack of effectiveness, and even outright failure. On the other hand, if a leader is able to mobilise the cognitive pressure of professionalism in favour of IT change, then an organisational culture of acceptance is possible (Prasad & Prasad, 1994).

**FUTURE TRENDS**

When considering the future of human factors in public information systems, it is useful to first recognise that an astonishingly rapid rate of technological change has been and is the norm for computing systems. Whole buildings and highly trained teams were once needed for a single computer; in contrast, more powerful systems can now be found on the top of the desk of the least-skilled person in an organisation. Given this rapid and dramatic pace of change and development, it is challenging to predict future directions.

Despite this dramatic pace which defies the imagination it is possible to observe that whatever technology may bring the core emphasis of human factors will remain unchanged. For any technological change to be effective, it must be readily understandable and usable by an average user. For example, while nearly instantaneous processing and presentation of data is becoming the norm, that data are of no value if confusing or overwhelming.

Any future system, regardless of any technological advance it may achieve, must include in its design consideration how humans comprehend, manage, and process information. For example, Miller (1956) reported that the average person is unable to remember more than seven pieces
of information at one time. A computing system that requires that a user remember eight pieces of information to successfully complete an operation will therefore have high rates of user error. It is this kind of simple, yet critical, understanding of human capability that must always be included in design.

Thus, technologies and designs of the future will always need to include recognition of human limitations in any interface design. Toolkits that contain information about human capabilities will make it possible for researchers and designers to consistently include consideration of the human user in any final product.

CONCLUSION

New technologies are often seen as an ultimate solution. While this is essentially correct, it is also necessary to recognise that improved technology does not change human capability. Computers and computing systems, as with public information systems, can only be as effective as users are capable. Designs that expect users to learn to perform beyond usual capabilities ignore the reality that humans are limited and unlike technology, cannot be improved or upgraded. For this reason it is essential that the concept of user-centred design be part of any public information system design process.

As implied, the introduction into IT system development of user-centred design procedures ensures that success stories associated with usability engineering will continue. Advanced development projects that examine psychological factors underlying consumer reactions and expectations include user-centred evaluations of prototypes and match human capabilities with system designs. These are most likely to result in public information systems that are accepted by both the public and governmental employees.

The inclusion of human factors into public information system development and design processes will, over the life of a system, reduce costs through the avoidance of the need to correct problems caused by errors or through difficulty in using the system. To successfully achieve this integration strong leadership that requires human factors elements in a system is necessary. Strategies such as specifying ISO 13407 (1999) as the standard during development and design are recommended as a way to be sure that human factors will be included.

REFERENCES


Cooper, A. (1999). The inmates are running the asylum. Indianapolis, IN: SAMS.


Human Factors in Public Information Systems


**KEY TERMS**

**Cognitive Engineering:** Understanding and predicting how changes in a task environment will influence task performance.

**Human-Computer Interaction (HCI):** The study of people, computer technology, and the ways these influence each other (Dix et al., 1998).

**Human Factors (or Ergonomics):** The scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimise human well-being and overall system performance (IEA, 2005).
**Interface:** A point in any human-computer system where information is exchanged. Interface designs that do not take into consideration human capabilities will likely have higher than desirable rates of human error.

**Usability:** The effectiveness, efficiency, and satisfaction with which specified users can achieve specified goals in particular environments.

**User-Centred Design:** During the design of systems or products the capabilities of end users (humans) are considered so that the final system or product is readily usable and designed so that the likelihood of error is minimised.

**User Interface:** The way that a human and computer system exchange information. Both hardware and software are part of the user interface. The screen, keyboard, and mouse are the most common hardware combination; any other hardware design that can be used to exchange information is also a user interface. The software arranges the information (e.g., sequence, quantity, manner requested, etc.). Human factors focuses on making certain that information is presented in a way that facilitates understanding.

*This work was previously published in Encyclopedia of Digital Government, edited by A. Anttiroiko and M. Malkia, pp. 940-946, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 1.6
Task Analysis at the Heart of Human–Computer Interaction

Dan Diaper
Middlesex University, UK

INTRODUCTION

The history of task analysis is nearly a century old, with its roots in the work of Gilbreth (1911) and Taylor (1912). Taylor’s scientific management provided the theoretical basis for production-line manufacturing. The ancient manufacturing approach using craft skill involved an individual, or a small group, undertaking, from start to finish, many different operations so as to produce a single or small number of manufactured objects. Indeed, the craftsperson often made his or her own tools with which to make end products. Of course, with the growth of civilisation came specialisation, so that the carpenter did not fell the trees or the potter actually dig the clay, but still each craft involved many different operations by each person. Scientific management’s novelty was the degree of specialisation it engendered: each person doing the same small number of things repeatedly.

Taylorism thus involved some large operation, subsequently called a task, that could be broken down into smaller operations, called subtasks. Task analysis came into being as the method that, according to Anderson, Carroll, Grudin, McGrew, and Scapin (1990), “refers to schemes for hierarchical decomposition of what people do.” The definition of a task remains a “classic and under-addressed problem” (Diaper, 1989b). Tasks have been differently defined with respect to their scope: from the very large and complex, such as document production (Wilson, Barnard, & MacLean, 1986), to the very small, for example, tasks that “may involve only one or two activities which take less than a second to complete, for example, moving a cursor” (Johnson & Johnson, 1987). Rather than trying to define what is a task by size, Diaper’s (1989b) alternative is borrowed from conversation analysis (Levinson, 1983). Diaper suggests that tasks always have well-defined starts and finishes, and clearly related activities in between. The advantage of such a definition is that it allows tasks to be interrupted or to be carried out in parallel.

Task analysis was always involved with the concept of work, and successful work is usually defined as achieving some goal. While initially applied to observable, physical work, as the field of ergonomics developed from World War II, the
task concept was applied more widely to cover all types of work that “refocused attention on the information processing aspect of tasks and the role of the human operator as a controller, planner, diagnostician and problem solver in complex systems” (Annett & Stanton, 1998). With some notable exceptions discussed below, tasks are still generally defined with people as the agents that perform work. For example, Annett and Stanton defined task analysis as “[m]ethods of collecting, classifying and interpreting data on human performance.”

BACKGROUND

Stanton (2004) suggests that “[s]implistically, most task analysis involves (1) identifying tasks, (2) collecting task data, (3) analyzing this data so that the tasks are understood, and then (4) producing a documented representation of the analyzed tasks (5) suitable for some engineering purpose.” While there are many similar such simplistic descriptions, Stanton’s five-item list provides an adequate description of the stages involved in task analysis, although the third and fourth are, in practice, usually combined. The following four subsections deal with them in more detail, but with two provisos. First, one should always start with Stanton’s final item of establishing the purpose of undertaking a task analysis. Second, an iterative approach is always desirable because how tasks are performed is complicated.

The Purpose of a Task Analysis

Task analysis has many applications that have nothing to do with computer systems. Even when used in HCI (human-computer interaction), however, task analysis can contribute to all the stages of the software-development life cycle. In addition, task analysis can make major contributions to other elements associated with software development, in particular the preparation of user-support systems such as manuals and help systems, and for training, which was the original application of hierarchical task analysis (HTA; Annett & Duncan, 1967; Annett, Duncan, Stammers, & Gray, 1971). HTA was the first method that attempted to model some of the psychology of people performing tasks.

Although infrequently documented, identifying the purposes for using task analysis in a software project must be the first step (Diaper, 1989a) because this will determine the task selection, the method to be used, the nature of the outputs, and the level of analysis detail necessary. The latter is vital because too much detailed data that does not subsequently contribute to a project will have been expensive to collect, and too high a level will require further iterations to allow more detailed analysis (Diaper, 1989b, 2004). Decomposition-orientated methods such as HTA partially overcome the level-of-detail problem, but at the expense of collecting more task data during analysis. Collecting task data is often an expensive business, and access to the relevant people is not always easy (Coronado & Casey, 2004; Degen & Pedell, 2004; Greenberg, 2004). Within a software-development life cycle, Diaper (2004) has suggested that one identify all the stages to which a task analysis will contribute and then make selections on the basis of where its contribution will be greatest.

Identifying Tasks

In the context of task scenarios, which Diaper (2002a, 2002b) describes as “low fidelity task simulations,” Carroll (2000) rightly points out that “there is an infinity of possible usage scenarios.” Thus, only a sample of tasks can be analysed. The tasks chosen will depend on the task analysis’ purpose. For new systems, one usually starts with typical tasks. For existing systems and well-developed prototypes, one is more likely to be concerned with complex and difficult tasks, and important and critical ones, and, when a system
is in use, tasks during which failures or problems have occurred. Wong (2004) describes his critical decision method as one way of dealing with the latter types of tasks.

Unless there are overriding constraints within a software project, then task analysts should expect, and accept, the need to be iterative and repeatedly select more tasks for analysis. Since the coverage of all possible tasks can rarely be complete, there is a need for a systematic task selection approach. There are two issues of coverage: first, the range of tasks selected, and second, the range of different ways that tasks may be carried out, both successfully and unsuccessfully.

One criticism of task analysis is that it requires extant tasks. On the other hand, all tasks subjected to task analysis are only simulations as, even when observed in situ, a Hiesenberg effect (Diaper, 1989b) can occur whereby the act of observation changes the task. Often, it is desirable to simulate tasks so that unusual, exceptional, and/or important task instances can be studied and, of course, when a new system or prototype is not available.

**Collecting Task Data**

There are many myths about task analysis (Diaper et al., 2003), and one of the most persistent involves the detailed observation of people performing tasks. Sometimes, task-analysis data do involve such observation, but they need not, and often it is inappropriate even with an existing system and experienced users.

Johnson, Diaper, and Long (1984; see also Diaper, 1989b, 2001) claim that one of the major strengths traditionally associated with task analysis is its capability to integrate different data types collected using different methods. The critical concept is that of fidelity. According to Diaper (2002a, 2002b), “fidelity, a close synonym is validity, is the degree of mapping that exists between the real world and the world modelled by the (task) simulation,” although as he says parenthetically, “N.B. slightly more accurately perhaps, from a solipsistic position, it is the mapping between one model of the assumed real world and another.”

At one end of the task-fidelity spectrum there is careful, detailed task observation, and at the other, when using scenarios of novel future systems, task data may exist only in task analysts’ imagination. Between, there is virtually every possible way of collecting data: by interviews, questionnaires, classification methods such as card sorting, ethnography, participative design, and so forth. Cordingley (1989) provides a reasonable summary of many such methods. The primary constraint on such methods is one of perspective, maintaining a focus on task performance. For example, Diaper (1990) describes the use of task-focused interviews as an appropriate source of data for a requirements analysis of a new generation of specialised computer systems that were some years away from development.

**Task Analysis and Task Representation**

The main representation used by virtually all task analysis methods is the activity list, although it goes by many other names such as a task protocol or interaction script. An activity list is a prose description of one or more tasks presented as a list that usually has a single action performed by an agent on each line. Each action on an activity-list line may involve one or more objects, either as the target of the action or as support for the action, that is, as a tool. An important component of an activity list should be the identification of triggers (Dix, Ramduny-Ellis, & Wilkinson, 2004). While most tasks do possess some sequences of activity list lines in which the successful completion of an action performed on one line triggers the next, there are many cases when some event, either physical or psychological, causes one of two or more possible alternatives to occur.
Diaper (2004) suggests that an activity list is sometimes sufficient to meet a task analysis’ purposes. He suggests that one of the main reasons for the plethora of task analysis methods is the volume of data represented in the activity list format, often tens, if not hundreds, of pages. As Benyon and Macaulay (2002) discuss, the role of task analysis methods applied to activity lists is not only to reduce the sheer amount of the data, but to allow the data to be abstracted to create a conceptual model for designers.

The two oldest and most widely cited task-analysis methods are HTA (Annett, 2003, 2004; Shepherd, 2001), and goals, operators, methods, and selection rules (GOMS; Card, Moran, & Newell, 1983; John & Kieras, 1996; Kieras, 2004). HTA is often misunderstood in that it produces, by top-down decomposition, a hierarchy of goals, and these are often confused with physical or other cognitive activities. HTA uses rules to allow the goal hierarchy to be traversed. Analyses such as HTA provide a basic analysis (Kieras) that can then be used by methods such as GOMS. While often perceived as too complicated, it is claimed that GOMS provides good predictive adequacy of both task times and errors.

There are between 20 and 200 task analysis methods depending on how one counts them. This presents a problem as different methods have different properties and are suitable for different purposes. An agreed taxonomy of methods for method selection is still unavailable. In Diaper and Stanton (2004a), there are half a dozen different taxonomies. Diaper (2004), rather depressingly, suggests, “in practice, people either choose a task analysis method with which they are familiar or they use something that looks like HTA.”

Limbourg and Vanderdonkt (2004) produced a taxonomy of nine task analysis methods, abridged in Table 1. The methods have been reorganized so that they increase in both complexity and expressiveness down the table. References and further descriptions can be found in Diaper and Stanton (2004a).

As can be seen from Table 1, there is no accepted terminology across task analysis methods. An exception, noted by Diaper and Stanton (2004b), is that of goals and their decomposition and generalisation.

A number of recent attempts have been made to classify tasks into a small number of subtasks (Carroll, 2000; Sutcliffe, 2003; Ormerod & Shepherd, 2004). The latter’s subgoal template (SGT) method, for example, classifies all information handling tasks into just four types: act, exchange, navigate, and monitor. Underneath this level, they have then identified 11 task elements. The general idea is to simplify analysis by allowing the easy identification of subtasks, which can sometimes be reused from previous analyses.

**TASK ANALYSIS AT THE HEART OF HUMAN-COMPUTER INTERACTION**

Diaper and Stanton (2004b) claim that “[t]oday, task analysis is a mess.” Introducing Diaper (2002c), Kilgour suggests that Diaper should consider the “rise, fall and renaissance of task analysis.” While Diaper argues that really there has been no such fall, Kilgour is right that there was a cultural shift within HCI in the 1990s away from explicitly referring to task analysis. Diaper’s oft-repeated argument has been that whatever it is called, analysing tasks has remained essential and at the heart of virtually all HCI work. Diaper (2002a, 2002b) comments, “It may well be that Carroll is correct if he believes that many in the software industry are disenchanted with task analysis…It may well be that the semantic legacy of the term “task analysis” is such that alternatives are now preferable.”
Task Analysis at the Heart of Human-Computer Interaction

Central to Diaper’s current definition of task analysis, and the primary reason why task analysis is at the heart of virtually all HCI work, is the concept of performance. His definition (Diaper, 2004; Diaper et al., 2003) is as follows:

Work is achieved by the work system making changes to the application domain. The application domain is that part of the assumed real world that is relevant to the functioning of the work system. A work system in HCI consists of one or more human and computer components and usually many other sorts of thing as well. Tasks are the means by which the work system changes the application domain. Goals are desired future states of the application domain that the work system should achieve by the tasks it carries out. The work system’s performance is deemed satisfactory as long as it continues to achieve its goals in the application domain. Task analysis is the study of how work is achieved by tasks.

Most models and representations used in software engineering and HCI are declarative; that is, they describe things and some of the relationships between things, but not the processes that transform things over time. For example, data-flow diagrams are atemporal and acausal and specify only that data may flow, but not when and under what circumstances. In contrast, it is essential, for all successful task-analytic approaches, that performance is modeled because tasks are about achieving work.

Based on Dowell and Long’s (1989) and Long’s (1997) general HCI design problem, Diaper’s (2004) systemic task analysis (STA) ap-

Table 1. An abridged classification of some task analysis methods (based on Limbourg & Vanderdonkt, 2004)

<table>
<thead>
<tr>
<th>Method</th>
<th>Origin</th>
<th>Planning</th>
<th>Operationalisation</th>
<th>Hierarchy Leaves</th>
<th>Operational Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTA</td>
<td>Cognitive analysis</td>
<td>Plans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOMS</td>
<td>Cognitive analysis</td>
<td>Operators</td>
<td>Methods &amp; selection rules</td>
<td>Unit tasks</td>
<td>Operators</td>
</tr>
<tr>
<td>MAD*</td>
<td>Psychology</td>
<td>Constructors</td>
<td>Pre- &amp; postconditions</td>
<td></td>
<td>Tasks</td>
</tr>
<tr>
<td>GTA</td>
<td>Computer-supported cooperative work</td>
<td>Constructors</td>
<td>Basic tasks</td>
<td>Actions</td>
<td>System operations</td>
</tr>
<tr>
<td>MUSE</td>
<td>Software engineering &amp; human factors</td>
<td>Goals &amp; constructors</td>
<td>Actions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TKS</td>
<td>Cognitive analysis &amp; software engineering</td>
<td>Plans &amp; constructors</td>
<td>Procedures</td>
<td>Actions</td>
<td></td>
</tr>
<tr>
<td>CTT</td>
<td>Software engineering</td>
<td>Operators</td>
<td>Scenarios</td>
<td>Basic tasks</td>
<td>Actions</td>
</tr>
<tr>
<td>Dianne+</td>
<td>Software engineering &amp; process control</td>
<td>Goals</td>
<td>Procedures</td>
<td>Operations</td>
<td></td>
</tr>
<tr>
<td>TOOD</td>
<td>Process control</td>
<td>Input/output transitions</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
proach emphasizes the performance of systems. While STA is offered as a method, it is more of an approach in that it deals with the basics of undertaking the early stages of a task analysis and then allows other analysis methods and their representations to be generated from its activity list output. The advantage of STA over most other task analysis methods is that it models systems and, particularly, the performance of the work system. STA allows the boundary definition of a work system to change during a task so that different constituent subtasks involve differently defined work systems, and these may also be differently defined for the same events, thus allowing alternative perspectives.

The novelty of STA’s view of work systems is threefold. First, as the agent of change that performs work, a work system in HCI applications is not usually anthropocentric, but a collection of things, only some of them human, that operate together to change the application domain. Second, it is the work system that possesses goals concerning the desired changes to the application domain rather than the goals being exclusively possessed by people. Third, STA is not monoteleological, insisting that work is never achieved to satisfy a single goal, but rather it states that there are always multiple goals that combine, trade off, and interact in subtle, complex ways.

STA’s modeling of complex work systems has recently been supported by Hollnagel (2003b) in cognitive task design (CTD); he claims that “cognition is not defined as a psychological process unique to humans, but as a characteristic of systems performance, namely the ability to maintain control. The focus of CTD is therefore the joint cognitive system, rather than the individual user.” Hollnagel’s formulation of CTD is more conservative than STA’s; for example, CTD is sometimes monoteleological when he refers to a single goal, and he restricts nonhuman goals to a limited number of things, albeit “a growing number of technological artefacts” capable of cognitive tasks. In STA, it is not some limited number of technological artefacts that possess goals and other cognitive properties, but the work system, which usually has both human and nonhuman components.

**FUTURE TRENDS**

While recognising the difficulty, perhaps impossibility, of reliably predicting the future, Diaper and Stanton (2004b) suggest that one can reasonably predict possible futures, plural. They propose that “[f]our clusters of simulated future scenarios for task analysis organized post hoc by whether an agreed theory, vocabulary, etc., for task analysis emerges and whether task analysis methods become more integrated in the future.” While not predicting which future or combination will occur, or when, they are however confident that “[p]eople will always be interested in task analysis, for task analysis is about the *performance* of work,” even though they admit that “[l]ess certain is whether it will be called task analysis in the future.”

Probably because of its long history, there is an undoubted need for the theoretical basics that underpin the task concept and task analysis to be revisited, as Diaper (2004) attempts to do for the development of STA. Diaper and Stanton (2004b) also suggest that some metamethod of task analysis needs to be developed and that more attention needs to be placed on a wide range of types of validation, theory, methods, and content, and also on methods’ predictive capability to support design and for other engineering purposes (Annett, 2002; Stanton, 2002; Stanton & Young, 1999). At least two other areas need to be addressed in the future: first, how work is defined, and second, the currently ubiquitous concept of goals.

Task analysis has always been concerned with the achievement of work. The work concept, however, has previously been primarily concerned with employment of some sort. What is needed, as Karat, Karat, and Vergo (2004) argue, is a broader definition of work. Their proposals are consistent...
with STA’s definition of work being about the work system changing the application domain. They persuasively argue for nonemployment application domains, for example, domestic ones. Thus, a home entertainment system, television, or video game, for example, could be components of a work system, and the goals to be achieved would be to induce pleasure, fun, or similar feelings in their users. That such application domains are psychological and internal to such work systems’ users, rather than the more traditional changes to things separate and external to some work system’s components, is also consistent with STA’s conceptualisation of task analysis.

Finally, Diaper and Stanton (2004b) broach, indeed they attempt to capsize, the concept of goals. They question whether the goals concept is necessary, either as what causes behavior or as an explanation for behavior, which they suggest, based on several decades of social psychological research, is actually usually post hoc; that is, people explain why they have behaved in some manner after the event with reference to one or more goals that they erroneously claim to have possessed prior to the behavior. Not only in all task analysis work, but in virtually every area of human endeavour, the concept of goals is used. Abandoning the concept as unnecessary and unhelpful is one that will continue to meet with fierce resistance since it seems to be a cornerstone of people’s understanding of their own psychology and, hence, their understanding of the world. On the other hand, academic researchers have a moral duty to question what may be widely held shibboleths. Currently, goal abandonment is undoubtedly a bridge too far for nearly everyone, which is why STA still uses the goals concept, but greater success, if not happiness, may result in some distant future if the concept is abandoned. At the least, it is time to question the truth and usefulness of the goals concept.

CONCLUSION

Two handbooks (although at about 700 pages each, neither is particularly handy) on task analysis have recently become available: Diaper and Stanton (2004a) and Hollnagel (2003a). Both are highly recommended and, while naturally the author prefers the former because of his personal involvement, he also prefers the Diaper and Stanton tome because it provides more introductory material, is better indexed and the chapters more thoroughly cross-referenced, comes with a CD-ROM of the entire book, and, in paperback, is substantially cheaper than Hollnagel’s book. No apology is made for citing the Diaper and Stanton book frequently in this article, or for the number of references below, although they are a fraction of the vast literature explicitly about task analysis. Moreover, as task analysis is at the heart of virtually all HCI because it is fundamentally about the performance of systems, then whether called task analysis or not, nearly all the published HCI literature is concerned in some way with the concept of tasks and their analysis.

REFERENCES


Task Analysis at the Heart of Human-Computer Interaction

(Eds.), *The handbook of task analysis for human-computer interaction* (pp. 603-619). Mahwah, NJ: Lawrence Erlbaum Associates.


Long, J. (1997). Research and the design of human-computer interactions or “What happened to validation?” In H. Thimbleby, B. O’Conaill, & P. Thomas (Eds.), *People and computers XII* (pp. 223-243). New York: Springer.


Task Analysis at the Heart of Human-Computer Interaction


**KEY TERMS**

**Activity List:** A prose description of a task or subtask divided into lines to represent separate task behaviors and that usually has only one main agent and one action per line.

**Application Domain:** That part of the assumed real world that is changed by a work system to achieve the work system’s goals.

**Goal:** A specification of the desired changes a work system attempts to achieve in an application domain.

**Performance:** The quality, with respect to both errors and time, of work.

**Subtask:** A discrete part of a task.

**Task:** The mechanism by which an application domain is changed by a work system to achieve the work system’s goals.

**Work:** The change to an application domain by a work system to achieve the work system’s goals.

**Work System:** That part of the assumed real world that attempts to change an application domain to achieve the work system’s goals.

This work was previously published in *Encyclopedia of Human Computer Interaction*, edited by C. Ghaoui, pp. 579-587, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
ABSTRACT

In this chapter the evaluation of human computer interaction (HCI) with mobile technologies is considered. The ISO 9241 notion of ‘context of use’ helps to define evaluation in terms of the ‘fitness-for-purpose’ of a given device to perform given tasks by given users in given environments. It is suggested that conventional notions of usability can be useful for considering some aspects of the design of displays and interaction devices, but that additional approaches are needed to fully understand the use of mobile technologies. These additional approaches involve dual-task studies in which the device is used whilst performing some other activity, and subjective evaluation on the impact of the technology on the person.

INTRODUCTION

This chapter assumes that ‘usability’ is not a feature of a product, that is, it does not make sense to call a product itself ‘usable’. Rather, usability is the consequence of a given user employing a given product to perform a given activity in a given environment. Holcomb and Tharp (1991) proposed a ‘model’ of interface usability, which is illustrated by Table 1. The definitions presented in Table 1 arose from consideration of the user interface of desk-based computers. However, it ought to be apparent that the majority of the components are defined in terms of an individual’s perceptions of features of the user interface.

The International Standards Organization has a number of standards relevant to human-computer interaction (Bevan, 2001). Current standards for mobile devices tend to focus on product attributes, for example, ISO 18021: Information Technology —User Interface for Mobiles (2001) provides interface specifications for Personal Digital Assistants. Other Standards have recognized the multifaceted nature of usability and have sought to encourage an approach that is similar to Quality Assessment (Earthey et al., 2001). Demonstrating compliance with the standards requires analysts to document their evaluation, demonstrating how it meets the objectives of the standard. The
Evaluating Mobile Human-Computer Interaction

The definition of usability offered by the International Standards Organization, that is, in ISO9241, part 11, is, “... the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.” (ISO9241-11, 1998).

The implications are that, first, usability is the consequence of a given user employing a given product to perform a given activity in a given environment (as stated) and, second, that it is possible to measure aspects of this relationship in terms of effectiveness, efficiency, and user satisfaction. It is important to note that these three aspects are inter-connected and that any evaluation activity ought to try to measure some aspect of each (Frøkjær et al., 2000).

**Defining Evaluation Targets**

If one is able to speak of measures, then it makes sense to be able to determine some criteria that indicate good or poor performance on these measures. Good et al. (1986) proposed that it is important to define both evaluation targets and metrics that relate to these targets. For example, in a study of conferencing systems, Whiteside et al. (1988) identified 10 attributes that they felt reflected the use of the conferencing system, for example, ranging from a fear of feeling foolish to a number of errors made during task performance. For each attribute, Whiteside et al. (1988) defined a method for collecting data about that attribute, for example, questionnaires, observation, and so forth, and then set performance limits relating to best, worst, and planned levels. A study of a wearable computer for paramedics (Baber et al., 1999) used this concept to produce Table 2.

In Table 2, three measures of performance were undertaken, that is, predictive modeling (using critical path analysis), user trials, and performance improvement arising from practice. In addition, three subjective evaluation methods were used. Table 2 shows how the system met (or exceeded) some of the target criteria but fell below the target

---

**Table 1. Holcomb and Tharp’s (1991) “model” of interface usability**

<table>
<thead>
<tr>
<th>Component</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional</td>
<td>Able to accomplish tasks for which software is intended</td>
</tr>
<tr>
<td></td>
<td>Perform tasks reliably and without errors</td>
</tr>
<tr>
<td>Consistent</td>
<td>Consistent key definitions</td>
</tr>
<tr>
<td></td>
<td>Show similar information at same place on screens</td>
</tr>
<tr>
<td></td>
<td>Uniform command syntax</td>
</tr>
<tr>
<td>Natural and Intuitive</td>
<td>Learnable through natural conceptual model</td>
</tr>
<tr>
<td></td>
<td>Familiar terms and natural language</td>
</tr>
<tr>
<td>Minimal memorization</td>
<td>Provide status information</td>
</tr>
<tr>
<td></td>
<td>Don’t require information entered once to be re-entered</td>
</tr>
<tr>
<td></td>
<td>Provide lists of choices and allow picking from the lists</td>
</tr>
<tr>
<td></td>
<td>Provide default values for input fields</td>
</tr>
<tr>
<td>Feedback</td>
<td>Prompt before destructive operations like DELETE</td>
</tr>
<tr>
<td></td>
<td>Show icons and other visual indicators</td>
</tr>
<tr>
<td></td>
<td>Immediate problem and error notification</td>
</tr>
<tr>
<td></td>
<td>Messages that provide specific instructions for action</td>
</tr>
<tr>
<td>User help</td>
<td>Online help system available</td>
</tr>
<tr>
<td></td>
<td>Informative, written documentation</td>
</tr>
<tr>
<td>User control</td>
<td>Ability to undo results of prior commands</td>
</tr>
<tr>
<td></td>
<td>Ability to re-order or cancel tasks</td>
</tr>
<tr>
<td></td>
<td>Allow operating system actions to be performed within the interface</td>
</tr>
</tbody>
</table>
Evaluating Mobile Human-Computer Interaction

Table 2. Defining evaluation targets

<table>
<thead>
<tr>
<th>Factors</th>
<th>Method</th>
<th>Metrics</th>
<th>Worst</th>
<th>Target</th>
<th>Best</th>
<th>Current</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>CPA</td>
<td>Time</td>
<td>-15%</td>
<td>0</td>
<td>+5%</td>
<td>-2%</td>
</tr>
<tr>
<td>Task</td>
<td>User trials</td>
<td>Time</td>
<td>-15%</td>
<td>0</td>
<td>+5%</td>
<td>-10%</td>
</tr>
<tr>
<td>Practice</td>
<td>1st vs. 3rd trial</td>
<td>% change</td>
<td>1st &gt; 3rd</td>
<td>3rd &gt; 1st</td>
<td>0</td>
<td>3rd &gt; 1st</td>
</tr>
<tr>
<td>Subjective evaluation</td>
<td>SUS¹</td>
<td>Scale: 0-100</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>SUMI²</td>
<td>Scale: 0-100</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Heuristics</td>
<td>Scale: 0-10</td>
<td>&lt;6</td>
<td>6</td>
<td>&gt;6</td>
<td>8</td>
</tr>
</tbody>
</table>

for time (although it is not within the ‘worst’ case range). One benefit of such a technique is to allow the design team to decide whether there is a need for more effort to refine the device, or whether, having met (some or all of) the requirements, the design process can be closed.

The ISO9241 notion of usability requires the concept of evaluation targets, for example, one could begin with a target of “66% of the specified users would be able to use the 10 main functions of product X after a 30 minute introduction.” Once this target has been met, the design team might want to increase one of the variables, for example, 85% of the specified users, or 20 main functions, or 15 minute introduction, or might want to sign-off that target.

Why Conduct Evaluation?

The concept of usability that is used in this chapter (and in ISO9241) implies that changing any one of the variables {user, activity, device, environment} can have an impact on usability. This implication points to the well-known assertion that an activity that a designer of the product might find easy to perform could prove problematic for a user who has had little or no previous experience of the product. It also points to potential issues relating to the usability of mobile technology, particularly through consideration of the environment. If we think about sending a text-message from a handheld device, such as a mobile telephone or a Blackberry™, the activities involved could be somewhat different while sitting on a train versus walking down a busy street. This change in environmental setting will have a marked effect on usability of the device. This does not necessarily result from the design of the device itself but rather from the interactions between design, use, and environment. As Johnson (1998) pointed out, “HCI methods, models and techniques will need to be reconsidered if they are to address the concerns of interaction on the move.” (Johnson, 1998). The question for this chapter, therefore, is how best to address the relationship between user, activity, product, and environment in order to evaluate the usability of mobile technology. Related to this question is how evaluation might capture and measure this relationship, and then what can designers do to improve usability. This latter point is particularly problematic if one assumes that design is about creating a product rather than about creating an interaction.

Before considering these questions, it is worth rehearsing why one might wish to conduct evaluation. Baber (2005) notes that the primary reason for conducting evaluation, in HCI, is to influence design (ideally, to improve the product). This implies that evaluation ought never to be a one-off activity to be conducted at the end of the design lifecycle in order to allow a design to be signed-off (Gould & Lewis, 1985; Johnson, 1992). Rather, it means the following:

1. Evaluation is a recursive activity that cuts across the entire design lifecycle, for exam-
Evaluating Mobile Human-Computer Interaction

...ple, software engineers will run versions of the code to debug and check; product designers will continually critique and refine their concepts. What is not always apparent is the manner in which these processes could (or indeed ought) be made formal and to result in something that can be communicated to other members of the design team.

2. Evaluation should be incorporated into as many stages of design as possible—this points to (i) but also raises that questions of recording and communicating the results of evaluation in a manner that can be beneficial to the design process.

3. Evaluation should be designed to maximize the impact of the evaluation of the design stage in which it is used—the suggestion is that, rather engaging in evaluation as a mandated exercise to allow sign-off between stages, it ought to be an activity that positively advances the design process.

4. Evaluation should guide and inform design activity—the results of any evaluation should be reported in a manner that can lead to change in the design and can be reported in a manner that is transparent and reliable.

A final point to note is that evaluation is a process of comparing the product against something else, for example, other products, design targets, requirements, standards. Thus, evaluation requires a referent model (Baber, 2005). It is naïve to believe that one can “evaluate” something in a vacuum, that is, to think that one can take a single product and “evaluate” it only in terms of itself. In many ways this is akin the concept of a control condition in experimental design; one might be able to measure performance, but without knowing what would constitute a baseline for the measure, it is not possible to determine whether it is good or bad.

Defining Referent Models

While it might be fairly clear as to why comparison requires a referent model, there is a problem for novel technologies. After all, the point of these technologies is to move beyond the conventional desk-bound personal computers and this will ultimately create new forms of interaction. However, the move to different technologies makes it hard to establish a sensible basis for evaluation. What is the referent model for mobile HCI?

A common form of mobile technology is the digital tour-guide that, knowing where the user is (using Global Positioning Satellite (GPS) to determine location) and what the user is doing, can provide up-to-the-minute information to help the user. There are few, if any, products that are like these concepts, so what constitutes a referent? At one level, this is simply because future HCI is attempting to develop approaches to interaction with technology for which there are no existing models. The answer to this question, the author suggests, comes from the assertion at the start of this chapter: usability is not a characteristic of the product, but the result of the interactions between user, product, activity, and environment. If we assume that tourists have a variety of strategies and artifacts that they currently use to find out where they are or to find out interesting information about a particular location, for example, maps, books, leaflets, other people. One could ground an initial evaluation of using the product to perform a given set of activities in comparison with existing practices. Conducting evaluation against other products in terms of a set of activities offers the analyst the following benefits:

1. The evaluation will cover a range of functions on the products. It is important to ensure that the comparison provides a fair and accurate view of the product. After all, it is not really the point of evaluation to just demonstrate the product X is better than product Y—partly because there are bound
to be occasions when products X and Y are similar, or where product Y is better than product X, and partly because simply knowing that X > Y tells us very little about how to improve X (or Y) or why X is superior.

2. The focus of the evaluation is less on product functioning than on user activity. This might appear, at first glance, to be tautological—surely product evaluation is about evaluating the product? This is, of course, true in a technical sense. However, HCI is about human-computer interaction, and the defining feature of this relationship is the interaction (rather than either human or computer). If one is concerned with technical evaluation then, perhaps some of the features to be included in a comparison table (like the one shown in Table 2) would be some of the technical features, for example, processor speed, RAM, memory, and so forth.

3. As the evaluation is concerned with user activity (as opposed to product functioning), the type of metrics that could be applied may well change. When comparing user activity on two or more products, it is important to decide what information is really being sought. Do we want to know only that X > Y? Or do we want to know that using product X or Y have differing effects on user activity?

In the field of mobile and wearable computers, much of the evaluation research has focused on comparing performance on a wearable computer with performance using other media. Thus, studies might compare performance using a wearable computer, say to perform a task that involves following instructions, and find that sometimes performance is superior in the paper condition (Siegel & Bauer, 1997; Baber et al., 1999) and sometimes it is superior in the wearable computer condition (Bass et al., 1995, 1997; Baber et al., 1998). This highlights the potential problem of comparing disparate technologies in an evaluation; it is not clear that any differences in performance are due to the experiment favoring one technology over another or whether there are other factors at play here. For example, a common observation is that people using the wearable computer tend to follow the instructions laid out on the display, whereas people using paper tend to adopt a more flexible approach (Siegel & Bauer, 1997; Baber et al., 1999). The notion that technology influences the ways in which people work is often taken as ‘commonsense’ by Human Factors engineers. However, the question of how and why such changes arise ought to have a far deeper impact on evaluation than is currently the case. As mentioned earlier, one way to deal with this problem is to focus on activities that people are performing using a variety of products. However, this will only cope with part of the problem. For instance, the electronic tour-guide given could be evaluated in comparison with other ways of performing activities, but this does not tell us whether any differences between the electronic tour-guide and the other products are due to the concept or to the realization of the concept or to changes in the activity arising from the use of the device. In other words, if we find that the electronic tour-guide performs less well than speaking to someone, is this because the tour-guide lacks information, or because it lacks clear presentation of information, or because it lacks speedy access to the information, or because it lacks flexibility of response, or because of some other reason (the evaluation could point to all of these, not to specific reasons).

At one level, the evaluation of mobile HCI calls for the application of current evaluation techniques. However, there are other aspects of future HCI that call for rethinking of evaluation. In other words, it might not be entirely appropriate to take methods that have proven useful for evaluating desktop HCI and apply these to future HCI. As Wilson and Nicholls (2002) point out in discussing the evaluation of virtual environments:
Evaluating Mobile Human-Computer Interaction

There are only a limited number of ways in which we can assess people’s performance. We can measure the outcome of what they have done, we can observe them doing it, we can measure the effects on them of doing it or we can ask them about either the behavior or its consequences. (Wilson & Nicholls, 2002)

The underlying assumption here is that human behavior is measurable in a finite number of ways. Combining this assertion with the need to study the relationship between user, activity, device, and environment, it becomes apparent that evaluation of user interaction with mobile activity can be reduced to a small number of requirements. Furthermore, the ISO 9241 notions of efficiency, effectiveness, and satisfaction point to the approaches outlined by Wilson and Nicholls (2002). For example, efficiency could be considered in terms of the amount of resource expended in order to achieve a goal (perhaps in terms of time to complete a task), and effectiveness could relate to the quality of this performance (perhaps in terms of the amount of activity completed or the quality of the outcome), and satisfaction would relate to the user perception of the activity (perhaps in terms of a judgment relating to their own performance or effort, perhaps relating to some aspect of using the device). What is required is not so much a battery of new measures, so much as an adaptation of existing approaches that pay particular attention to the relatively novel aspects of the environment and activity that pertain to mobile devices.

Making sense of human activity with mobile technology

The argument so far is that what needs to be evaluated is not simply the product, but the interaction between user, activity, device, and environment. This raises the question of what can be defined as appropriate forms of activity. The first issue for mobile technology is the assumption that it is to be used on the move, which raises two possibilities: (1) ‘on the move’ means physically moving, for example, walking, driving a car, traveling as a passenger; (2) ‘on the move’ means being in different places away from ‘normal’ office environments. One problem relating to both of these possibilities is the difficulty of collecting data in the field—there are problems arising from recording the data, managing the collection of data, and controlling experimental conditions that are far from trivial. However, if evaluation studies involve managing the interactions between user, activity, device, and environment, then it might not be possible to concentrate efforts on specific aspects of the interactions, for example, comparing the use of the device under different mobility conditions.

Interacting with mobile technology while walking

Consideration of interaction while moving immediately suggests that asking people to evaluate a product whilst sitting down in a laboratory might lead to different results than when using the product ‘on the move’. This is just what Kjeldskov and Stage (2004) demonstrated. Indeed, they found that having participants report usability problems while sitting down in the laboratory led to more usability problems being reported than when the participants performed the evaluation while walking. They suggested that this result might have arisen from different demands on attention—in the seated condition there was little distraction from the product and so participants were able to devote most of their attention to it, but in the walking conditions, attention needed to be divided between the device and the task of walking. This effect can be compounded by variation in other contextual factors, such as lighting levels and complexity of the path that one is following (Barnard et al., 2007).
It has been demonstrated that walking can impact cognitive tasks (Ebersbach et al., 1995), and so the use of a mobile device could be thought of in terms of a ‘dual-task’. A common methodological approach in Ergonomics/Human Factors involves asking participants to perform one task while attending to another, for example, tracking a line on a screen while also performing mental arithmetic. There are several reasons why this approach is useful, both in terms of developing theory of human performance and in terms of considering how combinations of tasks can be modified. In broad terms, the assumption is that the human ability to process information from several sources can be compromised under conditions of increasing complexity. Complexity might arise from the difficulty of one or both of the tasks, from the quality of the signals being attended to, from the amount of interference between the two tasks, and so forth. By measuring performance on the tasks under different levels of complexity, it is possible to judge the person’s ability to perform and the amount of interference that could occur.

Taking the dual-task paradigm as a starting point, one can consider many forms of mobile technology to be used not only in different places but also while the person is physically moving, for example, walking down a busy street or following a predefined route or walking on a treadmill. Thus, one approach to studying mobile technology from a dual-task perspective would involve measuring some aspect of walking and some aspect of using the technology. Barnard et al. (2005) compared reading tasks on a Personal Digital Assistant (PDA) while walking on a treadmill and walking along a defined path. They found a reduction in walking speed (by around 33%) compared to walking without performing the tasks on the device. This indicates that using the device leads to measurable changes in walking activity. They found no difference in comprehension between conditions (although it is often difficult to find measurable differences in comprehension in experiments that involve reading from screens, see Dillon, 1992), but they did find that word search took significantly longer when walking along the path than on the treadmill. This result suggests that participants walking the path had more need to divide their attention between the world and the device, and indeed, path following correlated with the use of scroll bars on the device, suggesting that more attention on the world led to more need to scroll the text to find one’s place while reading. What is interesting about this particular study is that it reports objective results on both primary (using the device) and secondary (walking under different conditions) tasks, and shows some interactions between the two. This shows that modifying the environment has a bearing on activity which, in turn, affects user performance (even with the same device requiring the same activity).

Using Mobile Technology while On-the-Move

Prototypical mobile technologies often address scenarios related to tourists because this emphasizes the need to move around an unfamiliar environment and the desire for information relating to the world around us, for example, routes to places, interesting information about landmarks, advice on traveling, and so forth. Considering the scenario from the perspective of usability, evaluation could allow the design team to agree on ‘benchmark’ levels of performance using existing practices, and to then consider what benefits might accrue from modifying those practices through the introduction of technology.

While the activity of walking can interact with the use of the mobile device, there are other aspects of use on-the-move that can also play a role. Duh et al. (2006) compared the evaluation of a mobile telephone, used to perform a set of activities, in the laboratory and on a Mass Rapid Transit (MRT) train in Singapore. The study showed that participants encountered significantly more problems in the train condition than in the laboratory. The authors relate the problem to five primary areas: ambient noise levels, movement of the train, issues relating to privacy, increase in effort needed to perform
Evaluating Mobile Human-Computer Interaction

the activity, additional stress, and nervousness. Of these factors, the main ones relate to aspects of the environment, viz. noise and movement of the train, and these, in turn, have a bearing on the ability of participants to complete the activity. In addition to the affect of movement on the performance of the users, the impact on the performance of the technology is equally important, for example, what happens when wireless networks do not cover the whole of the area and the user encounters ‘shadows’, or what happens when positioning systems have drift or inaccuracies. One approach might be to attempt to guarantee optimal delivery of service at all times by modifying the infrastructure rather than the device, for example, with boosters located in the environment. Another approach would be to provide ways of informing the user about accuracy of the data on which the system is working (Bell et al., 2006).

Devices can also be used while driving automobiles and there is evidence that interference between the activity of driving and the activity of using the device are more than simply physical (Boase et al., 1988; Brookhuis et al., 1991; Svenson & Patten, 2005; Wikman et al., 1998). This means that using a ‘hands-free’ kit will not eliminate all forms of interference. For example, Nunes and Recarte (2002) show that the more cognitively demanding a telephone conversation, the greater the reduction in the user’s ability to attend to the environment while driving. This research further highlights the problem of isolating the usability of the device itself from the interactions between user, activity, and environment.

**SUBJECTIVE EVALUATION OF TECHNOLOGY**

“[U]ltimately it is the users of a software system [or any product] who decide how easy its user interface is to manipulate…” (Holcomb & Tharp, 1991). Thus, one might feel that asking people about the product would be the obvious and most useful approach to take. However, there are several problems with this approach, for example, people might not always be able to articulate how they feel about the product (so the reports might be incomplete or inconsistent), people might use a variety of previous experiences as their referent models (so it might be difficult to generalize results across respondents), people might not be able to respond critically to the product (so there might be a ‘halo-effect’ with the participant responding to the novelty of the device rather than considering issues of usability). For these and other reasons, it is common practice to provide some structure to subjective evaluation, usually through some form of procedure or checklist. Furthermore, it would be suggested that subjective evaluation should be used as a secondary measure as far as practicable, with the primary focus on data collected from user trials.

**Subjective Response to the Device**

Participants could be asked to walk-through the performance of a given activity using the device, by explaining what they are doing and why. Monk et al. (1986) presented a detailed set of guidelines on how to use walk-through approaches to the evaluation in their Cooperative Evaluation method. The main aim of the approach is to capture problems that users experience when using a product.

In terms of checklists, a great deal of research effort from the late 1980s to the mid 1990s led to the development of a number of usability surveys. Some, like CUSI-Computer User Satisfaction Inventory (Kirakowski & Corbett, 1988) and QUIS-Questionnaire for User Interface Satisfaction (Chin et al., 1988), are designed to capture user response to an interface, particularly in terms of affective components (such as satisfaction). Others, like the checklist of Ravden and Johnson (1989) or SUS (Brooke, 1996), have been designed to cover both aspects of the interface and characteristics of usability. While these surveys are based on sound HCI principles, interpretation is left to the analyst who could lead to potential bias or misinterpreta-
tion. The SUMI checklist (Kirakowski, 1996) was developed using a rigorous approach to defining appropriate components of usability and presents results in terms of a comparison with a database of previous evaluations.

Subjective Responses to Using the Device To Perform an Activity

In addition to eliciting opinions from users regarding the device, researchers are also keen to obtain reactions of some of the consequences of using the device. By way of analogy, if we consider the virtual reality research community, we can see efforts to elicit reaction to either the physical effects of using virtual reality, for example, Cobb et al.’s (1999) Virtual Reality Induced Symptoms and Effects (VRISE) or the measurement of ‘presence’ (Slater et al., 1994; Witmer & Singer, 1998). In the domain of wearable computers, physical effects have been evaluated using self-report on a comfort rating scale (Knight et al., 2002).

In terms of performing an activity, researchers often make use of the NASA-TLX (Hart & Staveland, 1988) which measure subjective response to workload. The basic notion is that activities make different demands on people in terms of time pressure or mental effort, and can lead to different responses such as frustration or perceived level of performance. The NASA-TLX captures these responses and can be used to compare perceptions of users with combinations of different devices or activities.

DESIGNING AN EVALUATION PROTOCOL

Throughout this chapter, emphasis has been placed on the notion of ‘context of use’ and the concept of usability defined by ISO 9241, pt. 11. The relationship between these concepts is illustrated by Figure 1. In order to evaluate any item of technology, one needs to plan an appropriate campaign of evaluation—this means consideration of the evaluation from the start of the design process and performance of evaluation as often as practicable during the course of development. Assume that, whatever design process is being followed, there will be four primary phases: initial concept development, prototyping, specification and build. At each phase, the form and type of evaluation will change (depending on access to functionality on the product as much as anything else), but the basic considerations remain constant, that is, adequately defining context of use and applying appropriate usability metrics.

Figure 1. ISO9241 usability evaluation process
Before elaborating on Figure 1 as a process, it is worth re-emphasizing the point made earlier that usability evaluation always involves comparison with the product being considered against some referent model. The referent model could be other products, but is equally likely to be a set of design targets (see Table 2). In terms of comparison, a set of usability metrics can be applied. The ‘efficiency’ metric relates to the manner in which resources are applied during the activity in order to achieve the outcome; the ‘effectiveness’ metric relates to the completion of the outcome; the ‘satisfaction’ metric relates to the user’s response to performing the activity. Needless to say, all metrics apply to a given user performing a given activity in order to achieve a given goal in a given context of use with a given product. In terms of what to measure, each metric has several options. For the sake of brevity, in this chapter, one quantitative and one qualitative measure for each metric will be considered (the reader is encouraged to review ISO 9241, pt. 11 as a starting point for considering alternatives). For ‘efficiency’, a quantitative metric could be the number of mistakes a person made when using the product, and a qualitative metric could be a subjective workload (using the NASA-TLX mentioned); for ‘effectiveness’, a quantitative metric could be time to achieve the goal and a qualitative metric could be a subjective rating of performance; for ‘satisfaction’, a quantitative metric could be time spent using the device (over the course of several days) and a qualitative metric could be a self-report of how pleasant the product was to use. It should be apparent that the distinction between efficiency, effectiveness, and satisfaction is somewhat arbitrary, which is why it is important to make sure that all three metrics are applied during evaluation.

The idea that evaluation requires a ‘protocol’ is meant to imply that one ought to approach it in much the same way that one approaches the design of an experiment, that is, by defining independent variables, which are the goal, activity, and context of use, and by defining dependent variables, which are the usability metrics. The notion of the referent model also makes sense in terms of experimental design because the ‘hypothesis’ under test is that the outcome will be equal to or better than the referent model.

**Initial Concept Development**

During the ‘initial concept development’ phase, it is possible that one of the components of ‘context of use’ will dominate the others. For example, a designer might have an idea about how the product will function or how to perform a particular task or how to help a particular user. In order to explore this concept, designers make use of scenarios in various forms, for example, storyboarding, sketching, rich pictures, illustrative stories, and so forth. From Figure 1, it can be argued that a good scenario would include (as a minimum) some consideration of the type of person who would be likely to use the product, the tasks that the person would perform in order to achieve specific goals (as well as any other tasks that might need to be performed concurrently), the environment in which they might be performing these tasks, and the presence or use of other products to support this activity. In the domain of ‘traditional’ computer systems, the ‘environment’ can be assumed to be more or less constant, that is, the computer would be used on a desk in an office. In mobile computing, the ‘environment’ will have a significant impact on how the product will be used, as will the range of tasks that the person will be performing. It is this impact of the environment and the increasing range of concurrent tasks that makes evaluating mobile technology different from other computer applications. One way in which these aspects can be considered is to develop a scenario in which a person achieves the defined goal using no technology, another in which they use ‘contemporary’ technology and another in which they use the concept product. By storyboarding these different scenarios, the design team gets a feeling for the main benefits to be gained from using the product (and an
appreciation as to whether or not to pursue its development). During this stage, the usability metrics can be defined in terms of what measures can sensibly differentiate the product from any alternative ways of performing the task.

**Prototyping**

During ‘prototyping’ different versions of the product are developed and tested. The prototype need not be a fully-functioning product. Indeed, Nilsson et al. (2000) shows how very simple models can be used to elicit user responses and behaviors. Their study involved the development of a handheld device (the ‘pucketizer’ for use in water treatment plants and the initial studies had operators walking around the plant with a non-functioning object to simulate the device. From this experience, the design team went on to implement a functioning prototype based on an 8-bit microcontroller, wireless communications, and a host computer running a JAVA application). This work is interesting because it illustrates how embedding the evaluation process in the environment and incorporating representative end-users lead to insights for the design team. Taking this idea further, it is feasible for very early prototyping to be based on paper versions. For example, one might take the form factor of the intended device (say a piece of wood measuring 5” x 3” x ½”—which is approximately the size of Personal Digital Assistant) and then placing 3” x 2” paper ‘overlays’ to represent different screen states—change the ‘screens’ is then a matter of the user interacting with buttons on the ‘product’ and the evaluator making appropriate responses. Of course, this could be done just as easily using an application in WinCE (or through the use of a slideshow on the device), but the point is that initial concepts can be explored well before any code is written or any hardware built.

**Specification and Build**

‘Specification and build’ is the phase that one might traditionally associate with evaluation. Evaluation activity at this phase of the design process would be ‘summative’ (i.e., occur at the summation of the process), and would usually be used to confirm that the design was acceptable prior to committing to manufacture. At this stage, the main concerns regarding hardware and software would have been dealt with and so any usability evaluation that would call for significant change to hardware or software is likely to be ignored (unless the product is scrapped and the process started again, or unless these recommendations are filed for the next version of the product). However, usability evaluation can play an important role in this phase because it will form part of the acceptance testing of end-users and could, if positive, play a role in defining marketing activity or informing training requirements.

**CONCLUSION**

While the concept of usability as multi-faceted might seem straightforward, it raises difficult problems for the design team. The design team focuses its attention on the device, but the concept of usability used in this chapter implies that the device is only part of the equation and that other factors relating to the user and environment can play significant roles. The problem with this, of course, is that these factors lie outside the remit of the design team. One irony of this is that a well-designed device can ‘fail’ as the result of unanticipated activity, user characteristics, and environmental features.

The issue raised in this chapter is that evaluating mobile technology involves a clear appreciation of the concept of usability, in line with ISO standard definitions. The ISO9241 concept of usability emphasizes the need to clearly articulate the ‘context of use’ of the device, through consideration of user,
activity, device, and environment. This means that evaluation has to take account of the interactions between user, activity, device, and environment. What is essential is that evaluation is conducted in a way that ensures a good fit between the ‘context of use’ in the real-world and that simulated in the laboratory. This does not mean that one needs to include all aspects of the real-world in the laboratory but that one is able to reflect key variables and that the evaluation is designed to ensure a balanced comparison. It would be easy to ‘prove’ that a given device was superior to any other device simply by ensuring that the test favored the device in question. It is equally easy to ‘prove’ that evaluation in the ‘laboratory’ do not reflect performance in the ‘real-world’. However, such studies often reflect a limited grasp of adequate experimental design and, ultimately, a poor understanding of science. One is not ‘proving’ that a product is well designed through evaluation. Rather one is demonstrating ‘fitness-for-purpose’ under a well-defined context of use.

REFERENCES


Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. Inter-
national Journal of Human-Computer Studies, 60, 599-620.


**KEY TERMS**

**Context of Use:** The combination of user, task, product, and environment during the achievement of a desired goal

**Dual-Task:** The performance of two (or more) tasks at the same time. This could involve simultaneous performance or could involve some form of time-sharing between the tasks.

**Effectiveness:** The ability of a given user to employ a given product to achieve a desired goal in a given context of use

**Efficiency:** The optimal expenditure of resources by a given user in using a given product to achieve a desired goal in a given context of use

**Referent Model:** A product (or set of metrics) against which a given product can be compared
Satisfaction: The subjective response of a user to interacting with a product

Usability: “... the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.” (ISO9241-11, 1998)

ENDNOTES

1 SUS: Software Usability Scale (Brooke, 1996)
2 SUMI: Software Usability Metrics Inventory, Kirakowski and Corbett (1993)

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 731-744, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 1.8
An Overview of Multimodal Interaction Techniques and Applications

Marie-Luce Bourguet
Queen Mary University of London, UK

INTRODUCTION

Desktop multimedia (multimedia personal computers) dates from the early 1970s. At that time, the enabling force behind multimedia was the emergence of the new digital technologies in the form of digital text, sound, animation, photography, and, more recently, video. Nowadays, multimedia systems mostly are concerned with the compression and transmission of data over networks, large capacity and miniaturized storage devices, and quality of services; however, what fundamentally characterizes a multimedia application is that it does not understand the data (sound, graphics, video, etc.) that it manipulates. In contrast, intelligent multimedia systems at the crossing of the artificial intelligence and multimedia disciplines gradually have gained the ability to understand, interpret, and generate data with respect to content.

Multimodal interfaces are a class of intelligent multimedia systems that make use of multiple and natural means of communication (modalities), such as speech, handwriting, gestures, and gaze, to support human-machine interaction. More specifically, the term modality describes human perception on one of the three following perception channels: visual, auditive, and tactile. Multimodality qualifies interactions that comprise more than one modality on either the input (from the human to the machine) or the output (from the machine to the human) and the use of more than one device on either side (e.g., microphone, camera, display, keyboard, mouse, pen, track ball, data glove). Some of the technologies used for implementing multimodal interaction come from speech processing and computer vision; for example, speech recognition, gaze tracking, recognition of facial expressions and gestures, perception of sounds for localization purposes, lip movement analysis (to improve speech rec-
An Overview of Multimodal Interaction Techniques and Applications

In 1980, the put-that-there system (Bolt, 1980) was developed at the Massachusetts Institute of Technology and was one of the first multimodal systems. In this system, users simultaneously could speak and point at a large-screen graphics display surface in order to manipulate simple shapes. In the 1990s, multimodal interfaces started to depart from the rather simple speech-and-point paradigm to integrate more powerful modalities such as pen gestures and handwriting input (Vo, 1996) or haptic output. Currently, multimodal interfaces have started to understand 3D hand gestures, body postures, and facial expressions (Ko, 2003), thanks to recent progress in computer vision techniques.

BACKGROUND

In this section, we briefly review the different types of modality combinations, the user benefits brought by multimodality, and multimodal software architectures.

Combinations of Modalities

Multimodality does not consist in the mere juxtaposition of several modalities in the user interface; it enables the synergistic use of different combinations of modalities. Modality combinations can take several forms (e.g., redundancy and complementarity) and fulfill several roles (e.g., disambiguation, support, and modulation).

Two modalities are said to be redundant when they convey the same information. Redundancy is well illustrated by speech and lip movements. The redundancy of signals can be used to increase the accuracy of signal recognition and the overall robustness of the interaction (Duchnowski, 1994).

Two modalities are said to be complementary when each of them conveys only part of a message but their integration results in a complete message. Complementarity allows for increased flexibility and efficiency, because a user can select the modality of communication that is the most appropriate for a given type of information.

Mutual disambiguation occurs when the integration of ambiguous messages results in the resolution of the ambiguity. Let us imagine a user pointing at two overlapped figures on a screen, a circle and a square, while saying “the square.” The gesture is ambiguous because of the overlap of the figures, and the speech also may be ambiguous if there is more than one square visible on the screen. However, the integration of these two signals yields a perfectly unambiguous message.

Support describes the role taken by one modality to enhance another modality that is said to be dominant; for example, speech often is accompanied by hand gestures that simply support the speech production and help to smooth the communication process.

Finally, modulation occurs when a message that is conveyed by one modality alters the content of a message conveyed by another modality. A person’s facial expression, for example, can greatly alter the meaning of the words he or she pronounces.

User Benefits

It is widely recognized that multimodal interfaces, when carefully designed and implemented, have the potential to greatly improve human–computer interaction, because they can be more intuitive, natural, efficient, and robust.

Flexibility is obtained when users can use the modality of their choice, which presupposes that the different modalities are equivalent (i.e., they can convey the same information). Increased robustness can result from the integration of redundant, complementary, or disambiguating inputs. A good example is that of visual speech recognition, where audio signals and visual signals are combined to increase the accuracy of speech
recognition. Naturalness results from the fact that the types of modalities implemented are close to the ones used in human-human communication (i.e., speech, gestures, facial expressions, etc.).

**Software Architectures**

In order to enable modality combinations in the user interface, adapted software architectures are needed. There are two fundamental types of multimodal software architectures, depending on the types of modalities. In feature level architectures, the integration of modalities is performed during the recognition process, whereas in semantic level architectures, each modality is processed or recognized independently of the others (Figure 1).

Feature-level architectures generally are considered appropriate for tightly related and synchronized modalities, such as speech and lip movements (Duchnowski et al., 1994). In this type of architecture, connectionist models can be used for processing modalities because of their good performance as pattern classifiers and because they easily can integrate heterogeneous features. However, a truly multimodal connectionist approach is dependent on the availability of multimodal training data, and such data currently is not available.

When the interdependency between modalities implies complementarity or disambiguation (e.g., speech and gesture inputs), information typically is integrated into semantic-level architectures (Nigay et al., 1995). In this type of architecture, the main approach for modality integration is based on the use of data structures called *frames*. Frames are used to represent meaning and knowledge and to merge information that results from different modality streams.

**MAIN ISSUES IN MULTIMODAL INTERACTION**

**Designing Multimodal Interaction**

Recent developments in recognition-based interaction technologies (e.g., speech and gesture recognition) have opened a myriad of new possibilities for the design and implementation of multimodal interfaces. However, designing systems that take advantage of these new interaction techniques are difficult. Our lack of understanding of how different modes of interaction can be combined best into the user interface often leads to interface designs with poor usability. Most studies to understand natural integration of communication modes are found in the experimental psychology research literature, but they tend to qualitatively describe human-to-human communication modes. Very few attempts have been made so far to qualitatively or quantitatively describe multimodal human-computer interaction (Bourguet, 1998; Nigay, 1995; Oviatt, 1997). Much more work is still needed in this area.

*Figure 1. Multimodal software architectures*
Implementing Multimodality

Developers still face major technical challenges for the implementation of multimodality, as indeed, the multimodal dimension of a user interface raises numerous challenges that are not present in more traditional interfaces (Bourguet, 2004). These challenges include the need to process inputs from different and heterogeneous streams; the coordination and integration of several communication channels (input modalities) that operate in parallel (modality fusion); the partition of information sets across several output modalities for the generation of efficient multimodal presentations (modality fission); dealing with uncertainty and recognition errors; and implementing distributed interfaces over networks (e.g., when speech and gesture recognition are performed on different processors). There is a general lack of appropriate tools to guide the design and implementation of multimodal interfaces.

Bourguet (2003a, 2003b) has proposed a simple framework, based on the finite state machine formalism, for describing multimodal interaction designs and for combining sets of user inputs of different modalities. The proposed framework can help designers in reasoning about synchronization patterns problems and testing interaction robustness.

Uncertainty in Multimodal Interfaces

Natural modalities of interaction, such as speech and gestures, typically rely on recognition-based technologies that are inherently error prone. Speech recognition systems, for example, are sensitive to vocabulary size, quality of audio signal, and variability of voice parameters (Halverson, 1999). Signal and noise separation also remains a major challenge in speech recognition technology, as current systems are extremely sensitive to background noise and to the presence of more than one speaker. In addition, slight changes in voice quality (due, for example, to the speaker having a cold) can significantly affect the performance of a recognizer, even after the user has trained it.

Several possible user strategies to prevent or correct recognition errors have been uncovered. Oviatt (2000) shows that in order to avoid recognition errors, users tend to spontaneously select the input mode they recognize as being the most robust for a certain type of content (modality selection strategy). When recognition errors occur, Suhm (2001) suggests that users be willing to repeat their input at least once, after which they will tend to switch to another modality (modality switching strategy). Finally, Oviat (2000) reports cases of linguistic adaptation, where users choose to reformulate their speech in the belief that it can influence error resolution—a word may be substituted for another, or a simpler syntactic structure may be chosen. Overall, much more research is still needed to increase the robustness of recognition-based modalities.

APPLICATIONS

Two applications of multimodal interaction are described.

Augmented Reality

Augmented reality is a new form of multimodal interface in which the user interacts with real-world objects and, at the same time, is given supplementary visual information about these objects (e.g., via a head mounted display). This supplementary information is context-dependent (i.e., it is drawn from the real objects and fitted to them). The virtual world is intended to complement the real world on which it is overlaid. Augmented reality makes use of the latest computer vision techniques and sensor technologies, cameras, and head-mounted displays. It has been demonstrated, for example, in a prototype to enhance medical surgery (Dubois, 1999).
Tangible Interfaces

People are good at sensing and manipulating physical objects, but these skills seldom are used in human-computer interaction. Tangible interfaces are multimodal interfaces that exploit the tactile modalities by giving physical form to digital information (Ishii, 1997). They implement physical objects, surfaces, and textures as tangible embodiments of digital information. The tangible query interface, for example, proposes a new means for querying relational databases through the manipulation of physical tokens on a series of sliding racks.

FUTURE TRENDS

Ubiquitous Computing

Ubiquitous computing describes a world from which the personal computer has disappeared and has been replaced by a multitude of wireless, small computing devices embodied in everyday objects (e.g., watches, clothes, or refrigerators). The emergence of these new devices has brought new challenges for human-computer interaction. A fundamentally new class of modalities has emerged—the so-called passive modalities—that corresponds to information that is automatically captured by the multimodal interface without any voluntary action from the user. Passive modalities complement the active modalities such as voice command or pen gestures.

Compared with desktop computers, the screens of ubiquitous computing devices are small or non-existent; small keyboards and touch panels are hard to use when on the move, and processing powers are limited. In response to this interaction challenge, new modalities of interaction (e.g., non-speech sounds) (Brewster, 1998) have been proposed, and the multimodal interaction research community has started to adapt traditional multimodal interaction techniques to the constraints of ubiquitous computing devices (Branco, 2001; Schaefer, 2003; Schneider, 2001).

CONCLUSION

Multimodal interfaces are a class of intelligent multimedia systems that extends the sensory-motor capabilities of computer systems to better match the natural communication means of human beings. As recognition-based technologies such as speech recognition and computer vision techniques continue to improve, multimodal interaction should become widespread and eventually may replace traditional styles of human-computer interaction (e.g., keyboard and mice). However, much research still is needed to better understand users’ multimodal behaviors in order to help designers and developers to build natural and robust multimodal interfaces. In particular, ubiquitous computing is a new important trend in computing that will necessitate the design of innovative and robust multimodal interfaces that will allow users to interact naturally with a multitude of embedded and invisible computing devices.

REFERENCES


An Overview of Multimodal Interaction Techniques and Applications


KEY TERMS

Active Modality: Modality voluntarily and consciously used by users to issue a command to the computer; for example, a voice command or a pen gesture.

Feature-Level Architecture: In this type of architecture, modality fusion operates at a low level of modality processing. The recognition process in one modality can influence the recognition process in another modality. Feature-level architectures generally are considered appropriate for tightly related and synchronized modalities, such as speech and lip movements.

Haptic Output: Devices that produce a tactile or force output. Nearly all devices with tactile output have been developed for graphical or robotic applications.

Modality Fission: The partition of information sets across several modality outputs for the generation of efficient multimodal presentations.

Modality Fusion: Integration of several modality inputs in the multimodal architecture to reconstruct a user’s command.

Mutual Disambiguation: The phenomenon in which an input signal in one modality allows recovery from recognition error or ambiguity in a second signal in a different modality is called mutual disambiguation of input modes.

Passive Modality: Information that is captured automatically by the multimodal interface; for example, to track a user’s location via a microphone, a camera, or data sensors.

Semantic-Level Architecture: In semantic level architectures, modalities are integrated at higher levels of processing. Speech and gestures, for example, are recognized in parallel and independently. The results are stored in meaning representations that then are fused by the multimodal integration component.

Visual Speech Recognition: Computer vision techniques are used to extract information about the lips’ shape. This information is compared with information extracted from the speech acoustic signal to determine the most probable speech recognition output.

This work was previously published in Encyclopedia of Human Computer Interaction, edited by C. Ghaoui, pp. 451-456, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 1.9
Multimodal and Federated Interaction

Frankie James
SAP Research, USA

Rama Gurram
SAP Research, USA

ABSTRACT
This chapter introduces the concepts of multimodal and federated interaction. Because multimodality means, simply, the combination of multiple modalities (or types of input and output), the authors first introduce some of the various modalities available for computer interaction. The chapter then discusses how multimodality can be used both in desktop and mobile computing environments. The goal of the chapter is to familiarize scholars and researchers with the range of topics covered under the heading “multimodality” and suggest new areas of research around the combination of modalities, as well as the combination of mobile and stationary computing devices to improve usability.

THE BASICS OF MULTIMODALITY
As was discussed in the introduction to “Liability,” interaction with a computer (or computer-based device) can take place using a variety of different forms or modalities. On the input side, information can be transferred from a human operator to the computer via keyboards, keypads, touch screens, mice, joysticks, spoken language, or even gesture and motion sensors. Information can be output through visual displays (large and small), audio displays (including spoken text and non-speech sounds; see also “Mobile Speech Recognition”), and tactile displays (such as Braille or raised-line displays), as well as more exotic forms, such as force-feedback (haptic) joysticks and mice, and olfactory (“aroma-based”) displays.

Each of the input and output channels mentioned above have their own benefits and limita-
Multimodal and Federated Interaction

The purpose of this chapter is to discuss the concepts of multimodal interaction, where two or more modalities (such as vision or audio) are combined. Multimodal interaction is frequently used to compensate for limitations in one interaction modality by providing a second one. For example, the limited visual display capabilities of a mobile device can be augmented by providing audio output, or speech input can be provided to a user with limited typing ability to increase data entry speed. Multiple modalities can also be combined within a single input (or output) to increase efficiency; a seminal example here is the combination of a spoken action command (e.g., “color this red”) with mouse or touch-screen selection of the object to be acted upon. There are, of course, many other reasons for using multimodality, which will be discussed later in this chapter.

We will begin in this introduction with a discussion of the different forms of multimodality, as well as the different purposes to which multimodal interactions can be applied. In the next two sections, we address the use of multimodality for desktop applications and on mobile devices, where the reasons and methods for using multimodality can be quite different. We then discuss the concept of device federation, where multiple devices (each with their own available modalities) can be combined within a single interaction. Finally, we conclude with the chapter summary.

DIFFERENT FORMS

The basic definition of multimodality is the use of more than one modality within a single interface. The availability of both keyboard and voice input is one of the most common examples of multimodality, as is the use of both visual (text or graphical) and audio output. Most of the five classical human senses (sight, hearing, touch, smell, and taste) can be used for both the input and output sides. Each sense allows a broad range of possibilities.

Table 1 gives a brief list of the types of input and output that are associated with the senses. The most common use of the sense of sight is in the visual presentation (output) of information on small and large displays. Sight can also be used for input: eye tracking can be used for selection or to gauge interest in a particular area of a screen, and retinal scanning can be used to identify the user.

Input options based on the sense of hearing include speech, for entering text or giving commands, speaker identification (to identify or authenticate the user), and even humming (Ghias et al., 1995). Audio output can be used for presenting written text (using text-to-speech), recorded audio files, document and interface structures, and sonifications of graph data [see the chapter “Mobile Speech Recognition” and James (1998) for an overview]. Speech input and audio-based output are useful in a variety of contexts, including mobile and vehicle-based scenarios, as well as accessibility.

The sense of touch is already commonly found in computer inputs today, through the use of keyboards, pointing devices, and touch screens. In addition to detecting simply that a key, button, or screen area has been clicked or touched, more advanced devices (such as game controllers and track pads) can also detect the amount of pressure exerted by the user. Handwriting and gesture are also gaining in popularity within certain contexts, along with the use of fingerprints for user iden-
Multimodal and Federated Interaction

In the near future, we can expect to see heat and motion-sensing inputs, possibly for tracking people and objects. Tactile and haptic displays are also used for output, especially to make computers accessible to disabled users. Tactile displays include Braille and raised-line displays, which can be felt with the fingertips or hands to recognize shapes and text. Haptic displays use force feedback, generally to the hands and fingers, to present textures and boundaries that represent elements within the user interaction.

Smell and taste have spawned the creation of some new output possibilities, including the intriguing, if somewhat irreverent, Edible Bits (Maynes-Aminzade, 2005) gustatory display. The applicability of these senses to human-computer interaction, however, appears to be limited to specialized applications and artistic contexts.

Multimodal applications differ in which modalities and techniques they choose to combine, as well as in the ways that the modalities are used together. Some applications use modalities in a redundant way, so that input can be given in any of the available modalities, and the same output is rendered simultaneously in multiple modalities. Other applications use the modalities to complement one another, combining multiple inputs or outputs to produce a single command or piece of output (also called multimodal fusion). The choice of whether to use redundant or complementary modalities is dependent on the goals, context of use, and target user group for the application. Examples and use cases for both of these methods are described below.

**DIFFERENT PURPOSES**

Applications are designed to use multimodal interactions for a wide variety of reasons. Having several modalities available can allow designers to create applications that are easier or more natural to use, or that can be used by more users within more contexts. Table 2 shows the different purposes of using multimodality, categorized by

<table>
<thead>
<tr>
<th>Sense</th>
<th>Input Types</th>
<th>Output Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sight</td>
<td>Eye tracking</td>
<td>Visual displays (small and large) of graphics and text</td>
</tr>
<tr>
<td></td>
<td>Retinal scan</td>
<td></td>
</tr>
<tr>
<td>Hearing</td>
<td>Speech</td>
<td>Text-to-speech</td>
</tr>
<tr>
<td></td>
<td>Speaker identification</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Humming</td>
<td></td>
</tr>
<tr>
<td>Touch/Gesture</td>
<td>Keyboard</td>
<td>Tactile display</td>
</tr>
<tr>
<td></td>
<td>Mouse/pointing device</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Handwriting</td>
<td>Braille display</td>
</tr>
<tr>
<td></td>
<td>Gesture and motion</td>
<td>Haptic/force feedback display</td>
</tr>
<tr>
<td></td>
<td>Pressure sensing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heat sensing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fingerprint recognition</td>
<td></td>
</tr>
<tr>
<td>Smell</td>
<td>Chemical odor-detecting sensors</td>
<td>Olfactory output</td>
</tr>
<tr>
<td>Taste</td>
<td>(currently none)</td>
<td>Gustatory output</td>
</tr>
</tbody>
</table>
Multimodal and Federated Interaction

Table 2. Purposes for using multimodality

<table>
<thead>
<tr>
<th>Type of Multimodality</th>
<th>Purpose</th>
</tr>
</thead>
</table>
| Complementary Modalities (Multimodal Fusion) | o natural interaction; good mapping to task  
| | o avoid “clutter”/overload (output)  
| | o disambiguation (input) – non-mobile |
| Replacement Modalities | o natural interaction; good mapping to task  
| | o reduce screen clutter  
| | o compensate for limited input (or output) capabilities – mobile |
| Redundant Modalities | o accessibility  
| | o improve efficiency  
| | o compensate for hands- or eyes-busy tasks – mobile  
| | o compensate for environment – mobile |

the way the modalities are combined in the interaction. In many cases, the purposes are valid both for mobile and non-mobile (i.e., desktop) devices; in other cases, there is special value in using multimodality for specific device categories.

The most common usage of complementary modalities (or multimodal fusion) is to attempt to make the interaction more natural, where modalities are chosen that are the best fit to different parts of the task. For example, pointing and other gestures are appropriate for selection, while voice may be more efficient for entering text. Complementary modalities can also be used to spread the input or output across different modalities, reducing the load on any one channel. One example here is on the output side, where the most important or urgent information is presented in audio while supporting information is presented visually. The benefit is that the user can get both the key and supporting information simultaneously by listening to the audio and attending to the visuals at the same time.

Replacement modalities, as the name implies, replace one modality for another. Replacement can produce a more natural interaction, when, for example, keyboard entry of a password is replaced by speaker identification in a system that already allows voice input. On a mobile device, replacement can also be used to compensate for the limited capabilities of the device—long portions of text output are difficult to read and scroll through on a small screen, and may be easier to comprehend using an audio presentation.

Redundant modalities provide users with the ability to choose from different interaction options. By allowing both voice and keyboard input, for example, an application is more accessible to users who can’t type well or who have a physical disability. Redundant output, presented both visually and auditorially, can benefit not only users with visual disabilities, but also users performing tasks requiring their visual attention to be focused elsewhere (for example, while driving). The remainder of this chapter presents these purposes in more detail within the context of example applications and research prototypes.

MULTIMODALITY ON THE DESKTOP

Because desktop computers are still the primary computing environment for most users, we begin our discussion of multimodality on this platform. People frequently see multimodality as unnecessary or even inadvisable on desktop computers—“I don’t want to talk to my computer, or have it talk back to me, in my office…someone might hear some of my private information!” This section will show that there are many cases...
where the uses of multimodal input or output (even speech and audio) are indeed very useful tools for desktop interaction.

In desktop settings, there are several uses of multimodality:

1. To make the interaction more natural, by mapping information to be input or output to the most appropriate modality;
2. To provide better and more usable security;
3. To increase accessibility for disabled users.

This section discusses each of these uses, and concludes with the presentation of an area of possible research around the use of multimodal output in desktop computing.

**Using Multimodality to Make Interaction More Natural**

Standard desktop systems use keyboard and mouse for input, with visual displays and some (limited) audio output. For many interactions, these tools and modalities are appropriate; for example, selecting an icon on a desktop or some text in a document is straightforward using a mouse or other pointing device. However, other interactions involving pointing devices require far more effort on the part of the user, when the mapping requires users to search long menus or complicated tool bars for desired commands (for example, when specifying the attributes of a shape or icon placed on a canvas). Keyboards, while useful for entering text into a document, are not as useful for issuing commands or entering passwords, where arcane syntax and structures must be remembered. By providing more input modalities to a desktop computer user, interactions can be optimized to be more natural, and may also reduce the amount of visual clutter on the display.

Speech is perhaps the most commonly available input modality after keyboards and pointing devices (although as Oviatt [1999] points out, it should by no means be considered the primary information carrying modality within a multimodal system). Speech recognition systems use grammars to define the commands or phrases that can be accepted by the system. Speech recognition grammars can range from types found in standard computer theory (regular grammars, context-free grammars) to statistical grammars (e.g., n-gram grammars, where the system uses the preceding n-1 words spoken by the user to help identify the n-th word).

**Replacement Modalities**

Within the desktop context, speech input can be used to make an interaction more natural and also more visually appealing, by eliminating the need to display complicated tool bars and menus. A striking example of this is the CommandTalk system, which enabled users to create, control, and modify battlefield missions and control the map display of the ModSAF battlefield system using spoken language and mouse input (Moore et al., 1997). Through the addition of the CommandTalk speech interface, the researchers were able to reduce the screen clutter of ModSAF, eliminating almost all of the tool bars that had previously been required for specifying details of checkpoints and ordinances, controlling the display, and specifying mission movements.

The architecture of CommandTalk included many natural language processing components to create an intuitive command language. The speech recognition grammar was designed to allow the same commands and object referents as would be used in a human-to-human interaction. Complications regarding the resolution of noun phrases (“the M1 platoon”) to specific units shown on the display, and the mapping of predicates to interface actions (“move,” which can map to one of two different commands depending on
the type of platoon to be moved) were handled automatically and made transparent to the user. By designing the grammar using a corpus of sample human-to-human dialogues, the interaction is made extremely natural for users who already understand the domain, and the system can be used without much initial training.

In CommandTalk, the spoken language interface provides an alternative means of input. CommandTalk also changed the visual display of the underlying application, although this is not strictly necessary when creating a multimodal version of an existing application. The available input modality, speech, was in this case used as a replacement modality. Replacing one modality with another should, however, be done with caution. The spoken language input for CommandTalk was designed to be natural for expert users in the domain, based on sample dialogues. Non-expert users who are unfamiliar with the domain will require some time to learn the command structure that is used for the task. Toolbars and buttons can act as visible reminders of the available commands—these reminders allow users (especially non-experts) to build up their understanding of the domain using recognition memory. Removing the visible reminders forces users to rely on recall memory (Lodding, 1982).

In multimodal systems designed for non-expert users, it may be more appropriate in many cases to add a new modality without removing the old; in this case, the speech input would have been a redundant modality.

Complementary Modalities

Input modalities can also be used in a complementary way, where the inputs from several modalities are fused to produce a single input command. This is commonly known as multimodal fusion. Multimodal fusion allows users to mix and match modalities as they choose, and the system attempts to find a coherent meaning for the set of inputs. For example, when speech and gesture input are combined, spoken commands can be used to specify actions, while mouse or pen gestures specify the objects or action details. This can be much more efficient or more usable than using the mouse to both select the object and the command, since the two parts can be done simultaneously and because the user does not have to search the visual menus or tool bars for the command. Combining voice and a pointing device is also more natural than using voice alone, since object selection and many other tasks (such as resizing and moving) can be tedious and error-prone to accomplish using voice alone (see, for example, the Voice-Enabled Portal project, described later) (Grasso et al., 1998; Tyfa & Howes, 2000).

Many multimodal fusion systems focus on the use of speech with other modalities. One of the first systems to combine voice and gesture was the “Put-That-There” system (Bolt, 1980). The system focused on the basic concept of combining voice and gesture into what Bolt called a “concerted, natural user modality.” The system allowed users to place shapes on a large screen display, when seated facing the screen. Gesture sensors were used to determine the placement of the shape, while speech commands gave additional information, including the type of shape, color, and size. The gestures disambiguated deictic references (references whose interpretation is relative to the context of the utterance) in the spoken commands, such as “that” and “there.” As Oviatt (1999) points out, deictic commands tend to dominate multimodal input, since they are also widely used in human-to-human communication.

The combination of speech with gestures has most commonly been demonstrated in map-related tasks. Koons et al. (1993) describe a system that integrates speech, gesture, and eye gaze input to query objects on a map. Their focus is on the interpretation of multimodal commands from the three modalities, using reasoning to combine the modalities and determine the objects and actions requested. The Multimodal Maps application designed by Adam Cheyer and Luc Julia (Cheyer

...
Multimodal and Federated Interaction

& Julia, 1998) is another example of a system that fuses gesture and voice inputs to produce a natural, less error-prone application (their paper is also a good reference to other publications related to multimodal fusion). The application is a travel planning tool that shows locations of interest on a map display. Users can ask questions using either modality, or both in combination. For example, users can give the spoken command “show restaurants within one mile of this hotel,” specifying the hotel by circling it on the map.

Multimodal commands involving speech can be formed and used in a variety of ways, and systems built to understand them have to be able to handle these differences. For example, while some commands may be formed by actions that co-occur, such as a circling gesture or pointing action that occurs simultaneously with the speech, many commands that should be combined are temporally separated: the gesture or pointing action may occur either before or after the spoken command. It is extremely important to properly associate the actions, since the information given in the spoken command may be impossible to fully interpret without the gesture, and vice versa. Standard algorithms pay attention to the temporal ordering of actions in various modalities, evaluating whether two actions are related based on how long the “gap” is between the actions, whether the individual actions themselves can be well-formed commands, and other criteria. Even systems that are designed with multimodal fusion in mind must be able to support unimodal commands (independent gestures or speech), since users typically intermix unimodal and multimodal commands (Oviatt, 1999).

Multimodality and Security

Computer security and usability often find themselves in conflict. For example, computer security suggests that user passwords should contain digits, non-alphanumeric characters, be more than six characters long, and not listed in a standard dictionary. Usability findings suggest that passwords that conform to these rules will be difficult to remember, causing users to compromise security by writing the passwords down for future reference. New modalities, in the form of biometrics and other techniques, may provide a solution to this problem.

Biometric identification systems include speaker identification, fingerprint recognition, and retinal scanning. These systems rely on obtaining a unique identifier for a user that is based on who the user is. In addition to reducing the need for users to remember an arcane security code, biometric identification can make a system easier to use by combining the identification task with other interaction tasks. For example, voice identification could theoretically be done while the user is giving a voice interaction command. The user experience is that the extra identification step is being “skipped,” when in fact it is being processed in the background.

Because current biometric identification systems are not 100% accurate, identifiers based on what the user knows (similar to passwords) will likely continue to be part of computer security. Multimodal techniques in this category include picture recognition tasks, where users must identify a familiar picture from among a set of unfamiliar ones (Weinshall & Kirkpatrick, 2004). Other applications are beginning to emerge within the picture recognition space, such as the SiteKey feature of the Bank of America website, where users select a picture to be displayed after they have logged in to online banking. The display of the user-selected picture reassures users that the pages and information they are viewing indeed come from the bank. Interestingly, decades-old research (Bakst, 1988) suggests that the best computer security will require a combination of who the user is, what the user knows, and something the user carries (for example, a keycard). By definition, a system that includes all of these would have to be multimodal.
Multimodality and Accessibility

New, often redundant, modalities can be added to an application to simply provide an alternative means to enter or receive information. One of the major reasons for doing this is to support the goal of accessibility, where computer systems are made to be usable by people with disabilities. By providing speech input capabilities, for example, users with physical disabilities that affect their ability to type or handle a pointing device can have access to the same applications that non-disabled users do. Similarly, audio (both speech and non-speech) output can be provided in addition to a visual screen to make an application or system accessible to blind users.

Accessible applications can be created in several ways, depending on when in the design process the issue is considered and the function of the application. Table 3 gives an overview of approaches that can be taken to make an application accessible, and lists benefits and drawbacks for each approach.

If accessibility is considered early in the process and a design is created that incorporates multiple input and output options to support a wide range of users, this is an example of *universal design.* Although there is some argument about whether it is in fact possible to create a design that is universally usable (Mace, 1998), especially given the difference between input modes and the kinds of information that they can most easily express (Oviatt, 1999), universal design is frequently applied to systems that are to be used in a public context. Legislation regarding equal accessibility to information (such as the Americans with Disabilities Act and Section 508 of the Rehabilitation Act in the United States) encourage the designers of kiosks, automated teller machines (ATMs), and other public computing sources to take this approach.

Some applications are designed early on to be accessible, but with a focus on a specific group of users. For example, several commercial systems and research projects [such as pwWebSpeak, ASTeR (Raman, 1993), and AHA (James, 1998)] have addressed the issue of giving blind users better access to WWW-based content, taking different approaches to presenting structured documents in audio. One of the interesting things about this area of research is that the underlying WWW documents are not altered, but are instead interpreted differently by the new application to achieve accessibility.

Table 3. Approaches to accessibility

<table>
<thead>
<tr>
<th>Approach Taken</th>
<th>Example</th>
<th>Benefits</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create new app.</td>
<td><em>Kiosk using “universal design”</em></td>
<td>o accessible to all user groups</td>
<td>o requires design planning for all potential user groups</td>
</tr>
<tr>
<td></td>
<td></td>
<td>o avoid “traps” by considering accessibility early</td>
<td>o must be started at initial design phase</td>
</tr>
<tr>
<td>Create app. targeted to a</td>
<td>*Audio-based Web browser for blind/visually-</td>
<td>o increase accessibility of existing content</td>
<td>o new app. may not be accessible to all</td>
</tr>
<tr>
<td>group of disabled users</td>
<td>impaired users</td>
<td>o no changes required to existing app.</td>
<td>o collaboration hard for disabled/non-disabled users</td>
</tr>
<tr>
<td>Add new modality to</td>
<td><em>Mac OS X VoiceOver</em></td>
<td>o increase accessibility of existing content</td>
<td>o requires original manufacturer to decide to make product change</td>
</tr>
<tr>
<td>existing app.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Create “helper” app.</td>
<td><em>Screen readers, Voice-Enabled Portal</em></td>
<td>o original application unchanged</td>
<td>o need access to underlying app.</td>
</tr>
</tbody>
</table>
Accessibility can also be added to products later on, as in the case of voice-enabled portal. The voice-enabled portal (VEP) [originally called “Voice over Workplace” in James & Roelands (2002)] is another example of a speech interface that was developed as a redundant means of input for an application, in this case a business portal. The system was designed to support physically disabled users by adding speech capabilities to multiple applications within a portal, without requiring those applications to be changed. To do this, the VEP interface supported navigation and basic text entry based on the application’s syntax. Users selected items by speaking their names or by referring to elements by their type (e.g., “text box”). A major problem for this system was in disambiguating commands that could refer to multiple targets: if there are four text boxes on the screen, to which one does the command “text box” refer? Because the main goal was to provide access for physically disabled users, the solution chosen uses voice input exclusively: VEP used a predefined prioritization scheme to prioritize targets based on their location on the screen, choosing the target with the highest priority. If more than one target had equally high priorities, VEP numbered the targets using semi-transparent icons and required the user to specify the desired one.

Screen readers are another type of application where accessibility is added after an application has already been designed and built. Screen readers give blind users access to graphical interfaces using audio feedback and text-to-speech systems to read back text on menus, within open documents, and related to icons. These systems typically use access “hooks” provided by the applications and the operating system to find text, icons, and other graphical elements. For physically disabled users, third-party speech recognition systems (as well as those offered by the operating systems vendors themselves) also use the “hooks” to generate command grammars that allow users to control applications as well as enter text.

Accessibility and design for users with disabilities may seem like a specialized area of work, with a relatively small target population. However, as pointed out in Perry et al. (1997) and other sources, users can be handicapped, or unable to accomplish a task, merely because of the circumstances in which they try to perform it. So-called temporary disabilities, such as the inability of a user to direct his visual attention to a computer monitor, can be mitigated through the use of alternative modalities.

Multimodal Output and Transferring Tasks between Modalities

Much of this section has focused on the use of multimodal input in desktop settings. Multimodal output can also be used in desktop computing, especially for accessibility, as in the example of screen readers for blind users. Another potential use for multimodal output in a desktop setting is related to computing tasks that may be continued in a non-desktop setting, such as reading and reviewing documents while mobile.

Imagine the case where a sales representative is in her office reviewing her previous interactions with a client whom she will be visiting later that day. She reads through meeting minutes, e-mails, and other text-based documents to refresh her understanding of the situation. At the meeting that afternoon, the client says something that reminds the sales representative of a note in one of the documents she read, so she wants to go back to the document to get more details. Because the sales representative is out of her office, she is only carrying a mobile device, which does not have a large enough screen to display the document. The device is, however, capable of presenting the information via an audio interface. Will an audio review of information that had been previously only viewed visually be as helpful as a reminder as would a visual glance?

The intriguing question rising from this and similar scenarios is based on the concept of source
memory (Johnson et al., 1993), or the memory of the qualities of the information presented (e.g., was it spoken or read, who said it, was it new information, etc.). If information is presented in one modality, but then must be retrieved using a different modality, does this affect the user’s success rate, efficiency, and other retrieval factors? If research shows the answer to be yes, it may turn out to be beneficial to support additional output modalities (e.g., speech and audio) in a desktop computing environment. Information can initially be presented in several modalities simultaneously, so that later review or search in any of the modalities is made easier.

**MULTIMODALITY FOR MOBILE DEVICES**

Increasingly, computer usage is moving from desktop-based systems towards mobile devices, including personal digital assistants (PDAs) and mobile smart phones. New business and consumer applications for mobile devices are being created on an almost daily basis, to provide services and information to mobile users. However, the design of mobile applications presents many challenges from the standpoint of usability.

Mobile devices are limited by their very nature in terms of the user experience they can provide. In order to remain portable, mobile devices are constrained to a small form factor, which means that their input and output devices are also small. Tiny keypads and touch screens make input a challenge, especially when the user is walking or moving through the world. Small output screens are difficult to read and cannot present much information at one time. This section describes how multimodal input and output support can enhance the usability of mobile devices.

**Input Modalities for Mobile Devices**

Mobile devices come in several varieties, two of the most common being PDAs and mobile phones. The default input for these two device groups differs: mobile phones typically use a keypad (usually a 12-button telephone keypad, with perhaps a few additional function buttons), whereas PDAs (especially consumer-based models) as well as certain smart phones (for example, the Palm-based Handspring series) typically rely on handwriting input on a pressure-sensitive screen or small keyboards. It is interesting to note that although phones were originally designed to facilitate voice communication, interaction with mobile phone devices for tasks other than talking to another person (e.g., messaging, changing settings, and dialing) still rely heavily on non-speech input and output.

The small keypads used by most mobile phones make it difficult for the user to efficiently enter text and perform other input functions. Telephone keypad mappings requiring multi-tap or predictive text entry can be error-prone and add additional steps to the user interaction. QWERTY keyboards on mobile devices are so small that the buttons can be difficult to accurately select. Initial attempts to mitigate these problems were taken by the device manufacturers themselves: PDA designers offered external keyboards that could be attached to the devices. The problem with this solution, as well as related research into keyboards that can be projected on a tabletop, is that keyboards are difficult (if not impossible) to use while standing, walking, or moving. If the user is forced to sit and remain stationary to use the keyboard, it defeats much of the purpose of the mobile device.

Keyboard alternatives have been proposed for mobile and wearable device input. These include a variety of gesture-sensing devices, such as the FingeRing device (Fukumoto & Tonomura, 1997), which senses chord-style typing using a set of wearable rings. Chord typing is based on
particular combinations of fingers tapped on a surface simultaneously; for example, tapping a combination (or “chord”) of index, middle, and ring finger describes a single letter of input. FingeRing allows users to tap chords on any surface, including a leg or arm, making it more useful while walking or moving. Other chord keyboards, such as the Twiddler©, are small hand-held devices with strategically placed keys that can be selected in groups to form the chords. The limitations of chording, though, are a small input alphabet (for FingeRing: five fingers, with two positions each per chord, yields only 32 different chords) and a fairly high cost of learning effort.

On the PDA side, handwriting recognition via stylus has been available since the early days of the Apple Newton. Unfortunately, problems with the quality of the recognition using this system were at least partially responsible for the device’s demise. More recently, the Graffiti system offered with Palm devices has seen better success, with its stroke-based (rather than word-based) recognition. Although Graffiti (or, rather, the underlying Unistrokes alphabet) was specifically designed to minimize recognition errors, Shneiderman (2002) also attributes the success of Graffiti to the ability of users to more easily see the locality and cause of errors than with word-based recognition.

Stylus input, then, seems to be a fairly good input modality for small devices. However, stylus, as well as the small keyboards found on Blackberry-style devices, face limitations in many mobile settings. These include situations where the user’s hands are busy or encumbered by gloves, or where the user’s eyes are busy. Both the stroke-based recognition used by Graffiti and small keyboards can be a bit slow and rather cumbersome for entering large amounts of text. Other modalities, such as voice, can alleviate some of these problems.

Voice input is a common and well-accepted alternative to stylus or keypad input for mobile devices. (“A Conversation with Jordan Cohen,” 2006) Mobile phones are especially well adapted to voice input, given that they already contain a microphone to support the telephony functions of the device. Voice input can free up the user’s hands to perform the other functions necessary when mobile, such as carrying packages, opening doors, steering a car, or wearing gloves. Of course, voice input has its limitations in a mobile context, as well. In extremely noisy environments, recognition quality can deteriorate to the point that voice is unusable. Quieter environments may also be problematic for voice input, when it is inappropriate to talk or when it would be undesirable to be overheard. Technologically speaking, speech recognition via mobile phone often has a higher error rate than desktop-based recognition, both due to potential background noise and the quality of the device’s microphone. And, limitations in device memory can force speech recognition grammars to be quite small, which impacts usability.

Another way to improve user input to a mobile device is to eliminate or reduce the need for the input. With new capabilities to determine the location of mobile phones, such as global positioning systems (GPS) and cell towers, mobile devices can automatically gather information that would otherwise require human input. For example, if a user conducts a search for restaurants, he is likely to want to find a restaurant close by. With an application that is not context-aware, the search query needed for this task will be something like “restaurant Palo Alto.” If the device is able to detect the user’s location, the query from the user would only have to be “restaurant,” and the application could use the device’s location to filter the results.

Context-awareness can be used in even more clever ways, as well. The MICA (Multimodal Interaction in Context-Adaptive Systems) project used context awareness to sense actions by the user that could be used as inputs (Lorenz et al., 2005). The warehouse-based prototype designed in the MICA project can sense when the user was passing an item that needed to be picked from a
shelf and placed on the user’s trolley, and indicate this to the user. In addition, the system was able to recognize when the user was looking for, but unable to find, an object on a shelf and provide guidance to the exact location.

Finally, mobile device input can be augmented using technologies that sense movements (e.g., tilting or lateral motion) of the device itself (Hinckley et al., 2005). For example, a phone can detect when it has been picked up and placed by the user’s ear to automatically activate the speaker. Or, a device can detect that it has been tilted, to scroll down on a menu. Such interactions can be achieved by augmenting the device with rotational and gyroscopic sensors, or by new techniques that make use of mobile phone cameras (Wang & Canny, 2006).

Output Modalities for Mobile Devices

The small screens available on mobile devices are not really adequate to display all of the complex information that is requested by mobile users, such as maps, directions, and documents. The most common way of addressing this is through the use of audio output to augment the small-screen display. For example, many navigation systems show a small portion of a map view on the phone screen, with turn-by-turn directions presented using voice. This type of interaction also benefits mobile users by allowing them to focus more of their visual attention on locomotion tasks, such as walking and driving, which are common when using mobile devices. The next section goes into more detail on the use of audio output in divided attention scenarios.

Mobile devices can also make use of haptic feedback as an output modality. The canonical example here is a vibration alert on a mobile phone. These alerts do not require either visual or auditory attention, making them useful in very noisy or very quiet environments while the user is mobile. Haptic feedback can be used in mobile applications, such as navigation, to support the movement-based input described above (e.g., tilting and rotation). Haptic output can provide the user with more feedback and control during tilting input, for example, by providing some level of resistance to help the user’s actions stay in sync with the action taken on the visual display (Oakley & O’Modhrain, 2005). More experimental uses of haptics include the expression of presence by a remote participant to enhance social interaction [for example, Rovers & van Essen (2004) and Teh et al. (2006), where hand-holding, petting, or hugging by the remote participant is expressed to the user through vibration on the local device].

Designing for Divided Attention

As mentioned above, mobile devices are used in many contexts where the user is unable to devote his or her full attention to the mobile device interface. These contexts include device interaction while walking or driving: situations where the mobile device task is secondary, and the task of locomotion or navigation through the world is primary. The user’s attention must be divided between the primary and secondary task, and the mobile device interaction must be sensitive to the fact that it is secondary. This means that the interaction must be designed to be usable without requiring undue amounts of attention, and must be especially sensitive to the modalities used to avoid overloading any of the sensory channels.

Multimodality can be used effectively in divided attention situations, most frequently by providing information input and output capabilities on channels that are not as important to the primary task. For example, visual attention requirements can be reduced by providing redundant output using the audio channel. This also reduces the need for the user to be in close physical proximity with the device, since audio can be heard over a greater distance than small visual displays can be used. Prototypes for warehouse picking have been designed in this way: the MICA system, for example, provides redundant visual and audio
information so that the worker can move away from the visual display on the warehouse trolley to retrieve an item from the shelf, while still receiving information about the item to retrieve (Lorenz et al., 2005).

An important caveat, however, is that simply changing modality does not always produce an interaction that can be used effectively in divided-attention scenarios, especially vehicle-based scenarios. A study on speech-based interaction in a driving context showed that menu navigation produced competition for attentional resources with the spatial aspects of navigating the vehicle (Lee et al., 2001). The central processing theory of attention suggests that this is because central resources are being overloaded, even if none of the individual modalities are (Moray, 1999). Therefore, divided-attention interactions should also be careful about the sheer amount of information that is to be presented, perhaps by adapting to the current driving situation based on information about traffic conditions, driver stress, and so forth.

Finally, investigations of vehicle interfaces have shown that the style of information presentation has a strong impact on driving ability. Lee et al. (1999) found that drivers responded differently, in terms of response time and action taken, depending on several factors. These factors included not only modality, but also display location and message style (e.g., whether the system advised of a situation or gave explicit instructions for action). Other studies have also shown that the phrasing of messages, as well as the emotions expressed by the voice used for output, have an impact on driving performance and attention (Jonsson et al., 2004; Jonsson et al., 2005). Voice gender can also make a difference, especially for certain tasks such as navigation and feedback on driving performance (Nass & Braves, 2005).

**DEVICE FEDERATION**

An emerging area of research in human-computer interaction involves the combination of small, portable devices with ambient computing and interaction resources in the user’s environment. This concept, which is cropping up in industrial and academic research projects at various locations, is an attempt to balance the dual problems of portability and usability through a new model for mobile interaction, which we will call here **device federation**.

The idea of device federation is to augment small, portable devices such as smart phones (called personal devices) with ambient computing resources, such as large displays, printers, computers, PDAs, and keyboards. The personal device can be used to establish the user’s identity (see the discussion above related to improving security by combining something the user carries with what the user knows and who the user is), run applications, or connect to back-end databases and servers, while ambient resources are leveraged to provide usable input and output. Conversely, personal devices can be connected to external sensors that have minimal or no user interfaces of their own, allowing the user to view or manipulate data that would otherwise be hidden within the environment.

Device federation is a broad concept, which covers a wide range of federation types. An obvious example is federating large displays with small mobile devices. Other types include federating sensors and other resources with limited human interaction capabilities with mobile devices, federating portable user devices with audio output, and federating portable input devices with ambient computers to provide accessibility.

**Ad Hoc Wireless Federation**

Device federation promises the ability to augment the large amount of computing resources found in today’s portable devices using the ambient devices
Multimodal and Federated Interaction

found nearly everywhere in the Western world. More and more offices, coffee shops, airports, and other locations contain large displays, speakers, and keyboards that could be used to make up for the limited input and output capabilities of mobile devices. Within the context of business usage, goals include allowing users to federate keyboards or other input devices with a personal mobile device to enhance input, or allowing voice input through a mobile device to be recognized and sent to a federated large-screen display.

It is already possible today to connect input and output devices to mobile phones and PDAs. USB and other wired connections provide a standard interface for keyboards, mice, displays, and other devices to connect to personal devices in a fairly seamless manner. However, wired connections require device proximity, which can be problematic for some use cases, such as connecting a mobile phone to a large-screen display. It would be preferable to be able to connect to the display from anywhere in the room, without having to be right next to it. In addition, some phones may not have standard wired connection points and instead use proprietary ports and connectors. Therefore, device federation holds wireless connectivity as one of its main goals.

The second main goal of device federation is to allow ad hoc connectivity. Many common wireless standards, such as Bluetooth, use handshaking and pairing protocols that are far too cumbersome to be of value when a personal device is to be connected to a particular ambient resource only once or twice in its lifetime. These protocols are also not very user-friendly, especially for inexperienced users.

One concern regarding ad hoc wireless federation is, of course, security. Devices that can easily connect and disconnect can also be easily compromised. Unfriendly devices could be used to snoop on data connections, or to grab computing resources. However, the need for security must be balanced with usability, so that the benefits of federation are not compromised.

Input and Output Federation

One of the most basic scenarios for device federation is display federation. Mobile devices often have screens that are too small to adequately display information, especially if the information involves complicated visualizations, graphics, or large quantities of text. Output federation can solve this problem by connecting with a nearby display to present the information. Additionally, input mechanisms to large screen displays have been investigated by many research groups [see Bezerianos & Balakrishnan (2005) for a recent example]. Federation offers another alternative, by allowing input through a mobile device to control a large-screen display.

The implementation of display federation is not as future reaching as it may seem. Intel's Ubiquity project is currently prototyping a device called the “Personal Server,” which could be used for this purpose. The personal server is “a small lightweight computer with high-density data storage capability.” The primary purpose of the device is to store data and provide computing power that can be leveraged by other devices (such as a laptop, desktop, or projection displays) using a local wireless connection (Want et al., 2002). In fact, some of the original incarnations of the Personal Server have included no direct input and output capabilities: the very use of this device depends on display federation (as well as federation to input devices). WinCuts from Microsoft Research allows users to push shared areas of live screen content to other machines, which are controlled locally (Tan et al., 2004). PdaReach by June Fabrics displays the contents of a PDA screen on a computer connected via the PDA synchronization cable.

What remains to be investigated are usage scenarios and design for display federation. For example, should all content automatically be routed to a federated large display, or should users explicitly control the portions of the output to be
displayed? Such considerations likely depend on the applications involved, the security or privacy of the environment, and other factors.

The federation of input devices would allow users to federate the input devices of their choice with other resources. Users could, for example, federate a speech input device to a computer, thereby allowing multimodal input to the computer’s applications. This would not only be beneficial to mobile users, but was also proposed more than ten years ago as a solution to the accessibility problem (Perry et al., 1997). The basic idea was to provide users with a portable device that exactly met their needs in terms of input capabilities: low-mobility users could use single-switch or voice entry, blind users could use Braille entry devices, and so forth. This would be an example of creating a “helper application” to support disabled users (as discussed earlier in this chapter): the new element here is that the helper application would not need to reside on the same computer as the original application.

The Apple iPod is also a good example of input federation. The iPod is portable and allows self-contained functionality, but its input and output functionality increase when federated with other devices such as computers and car stereos. Federation with car stereos is especially interesting in regards to the distribution of I/O: some products federate the iPod to a car stereo’s speakers, handling input via the iPod itself, while others use the car’s output and input. User satisfaction with the different products can provide insight into federation for other consumer or business applications.

Because flexible connection and disconnection is a goal for device federation, it is possible that users may want to federate (or de-federate) within the course of performing a task. This again raises the possibility of using redundant outputs to facilitate task transfer, as described above.

Data Source Federation

Data sources, in the form of sensors, RFID tags, and other “smart items” are becoming more prevalent in today’s environment. Sensors and tags can be used to track objects and provide real-time information. Federated devices can provide new opportunities to access these nearly invisible information sources by federating them with other devices (such as PDAs or laptop computers) that can display their state, or even modify their behavior.

Federating with data sources can also be used in ubiquitous computing environments to provide support for context-aware applications. Knowing what data sources are currently in proximity to the user’s personal device can give a good sense of location, and data such as noise level, temperature, and other environmental factors can be used to modify the interaction. For example, if the noise level reported by nearby sensors is very high, a mobile device can make the decision to present information using the visual and haptic channels, rather than using audio.

Research projects that investigate software updates for embedded and networked devices point the way for data source federation. The OSGi Alliance is creating an open service delivery and management platform so that “software components can be installed, updated, or removed on the fly without having to disrupt the operation of the device.” Similarly, Sun Labs is researching “Small Programmable Object Technology,” or Sun SPOT®, which allows running Java programs to be moved between sensors. The flexibility inherent in these projects opens the possibility of federating devices to access data from nearby sensors or tags.

In data source federation, it is important to investigate the process of connecting and disconnecting devices. This topic of course also arises in input and output federation scenarios, but the sheer volume of sensors and tags available within a local area increases the magnitude of the problem.
If the system requires users to search for nearby devices and explicitly grant pairing rights, selecting a few sensors (or a nearby display) from a list of thousands of possible devices becomes overwhelming. On the other hand, totally automatic pairing with thousands of nearby devices will generally be undesirable, as the user is unlikely to want to interact with all of them and may be uncomfortable not knowing which devices are connecting at any given time.

**SOCIAL IMPACT OF VOICE TECHNOLOGIES**

A discussion of multimodality would be incomplete without paying some consideration to the social impact of voice technologies. As mentioned earlier in the chapter, speech is one of the most common input technologies after keyboards and pointing devices. However, communicating with computer devices using speech raises many issues regarding how people respond to technologies that engage in this very “human” activity. The “Media Equation,” as it was coined by Reeves & Nass, can be summarized by the idea that people treat interactions with computers and other media the same way that they treat interactions in real life (Reeves & Nass, 1996). Among other things, users react to and expect the same social phenomena in human-computer communication as in human-human communication, including following the rules of politeness and interpersonal difference. Nass & Braves (2005) extend these findings to state that humans are “Wired for Speech,” and have even stronger expectations about how systems that can speak or listen should behave.

In this section, we will reflect on one fundamental issue related to multimodal interactions that is implied by this work: the mismatch of modalities between input and output (James, 1998). When a device (or desktop application, for that matter) provides voice output, but does not allow voice input, this puts the user in a subordinate position to the computer, and may make the user feel out of control. As Brenda Laurel puts it:

“We...tend to expect symmetry between the input and output modalities of a system; that is, we expect that they are operating in the same sensory universe as the rest of the representation. If a computer talks to us, we want to be able to talk back, and vice versa.... In most systems, our side of the I/O equation is severely impoverished. The system can present images, sounds, movements, words, and possibly even speech, but we must act inside a straitjacket of menu items and mouse clicks. No wonder we often feel that computers are always in control—the system is holding all the cards! Working toward symmetry in input and output channels in human-computer activities can vastly improve our experience of engagement and agency.” (Laurel, 1993)

Control is not the only issue. Mismatching modality can also be perceived as impoliteness. The polite response to a letter is another letter; answering a letter with a phone call is less polite. Computer systems that require an input in a different modality than the modality used for output, according to Reeves & Nass (1996), force the user to make an impolite response.

Due to the factors discussed above, care should be taken when adding a modality to make sure that it is available both for input and output, if it is technically possible to do so. There is, however, every possibility that new generations of computer users who have been exposed to interfaces with modality mismatches since birth may not have the same control and politeness responses [changes in politeness behavior related to technology use have been shown to some extent in relation to mobile phones, for example (Ling, 2004)]. Although computer interfaces evolve at a much faster pace than humans, people who grow up using multimodal interfaces that may not conform to human-human social interaction standards may adapt their reactions to the computers and be more accepting of seemingly impolite behavior.
SUMMARY

This chapter has described some of the basic principles of multimodality. We began with a description of some of the ways that the human senses can be used to interact with a computer, and discussed the idea that modalities can be used redundantly, where more than one modality is used to either present or gather the same information, or complementarily, where information from more than one modality must be combined to produce the whole input or output message. Each of these methods can be used for either desktop or mobile interactions, although generally for different reasons.

Multimodal interfaces to desktop computers strive to provide natural mappings between the information to be input or output and the modality used. This more frequently leads to interfaces that use complementary modalities, or multimodal fusion. Desktop applications also use multimodality to support users with disabilities or to provide more usable security. Another interesting research area for multimodality is around the use of redundant output modalities to support users who move from desktop to mobile computers during the course of an interaction.

Mobile applications use multimodality to improve the input and output capabilities of the devices. Because mobile devices must be small to be portable, they often have very small visual displays and keypads. Voice and audio are obvious choices here, and have been used widely. Finally, because mobile devices are often used in contexts where the user is carrying out another (more primary) task, such as walking or driving, it is important to design interfaces that do not require too much of the user’s attention or overload any of the senses.

Next, we described an emerging area of human-computer interaction that seeks to combine the portability of mobile devices with the interaction capabilities of larger ambient devices, called device federation. Input and output federation can, and have already begun to, be used to support mobile users with both business and consumer applications. Federation with data sources, such as sensors and RFID tags, could also benefit business users and support context-awareness.

The final section of this chapter discussed an important consideration when using voice interactions, namely, the social implications. Based on continuing research from sociology, it is clear that people treat media in many of the same ways that they treat other people, and have many of the same expectations. This is especially true for voice-based systems, so care must be taken when using voice input and output.

FUTURE RESEARCH DIRECTIONS

The importance of multimodal interaction design will continue to increase as users move more and more towards interaction with mobile devices, and away from traditional desktop computing environments. Increasing wireless connectivity is shifting users toward an “always-on, always-available” mentality that will be supported by newer and better mobile devices.

The widespread use of mobile devices will require a focus on supporting work that is done within the context of other activities. This work has already begun to some extent, but there are open questions around the types of tasks users choose to engage in while busy doing other things and how to manage interruptions within an interaction. For example, in the vehicle context, it will be important to understand what kinds of tasks are of interest to users, and then create interfaces that will help users complete them safely and without creating an undue mental burden. Even without considering interruptions, it will be important to investigate the confirmation of actions taken by users whose attention is divided between interaction with a mobile device and interaction with the rest of the world.
The federated device concept discussed in this chapter certainly falls under the area of future research. Initial projects have shown that it should be possible to (wirelessly and seamlessly) interconnect mobile and ambient devices; what remains to be seen is how the interactions with device federations should be structured. What information should be shown on a personal display, as opposed to a public display? Should the personal device be used for input to a device federation, and if so, how should it operate? Will the personal device act as a remote control, or will users respond to it simply as one component of the larger system? These and other questions are important to understand as the world moves toward ubiquitous computing environments.

Finally, this chapter listed a wide range of modalities available for human-computer interactions. Some of these modalities are already part of many computer systems, such as voice and audio, while others are less widely used. Haptics and other modalities could be valuable additions to new interaction designs, but more research is required to understand where they would be most appropriate for everyday use.

REFERENCES


**ADDITIONAL READING**


ENDNOTES

a Deictic references can also include reference to times, such as “now” or “later” and people (such as “you” or “me”).

b http://www.bankofamerica.com/privacy/index.cfm?template=privacysecurolb

c http://www.apple.com/macosx/features/voiceover/

d See http://trace.wisc.edu/world/gen_ud.html for a general discussion of the concept of universal design.

e http://www.soundlinks.com/pwgen.htm

f QWERTY is the name for a standard U.S. keyboard layout, and comes from the first six letter keys in the top row starting from the left. See Wikipedia for a more detailed explanation (http://en.wikipedia.org/wiki/QWERTY).

g http://www.handykey.com/site/twiddler2.html

h http://www.intel.com/research/exploratory/personal_server.htm

i http://www.junefabrics.com/index.php

j http://www.osgi.org/


This work was previously published in Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises, edited by M. Mühlhäuser and I. Gurevych, pp. 487-507, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
INTRODUCTION

In 801, Harun Rashid offered Charlemagne a water clock, the like of which was inexistent in all of Europe at that time; the King’s court thought that a little devil was hidden inside the clock. In the 1930s, King Abdulaziz of Saudi Arabia had to convince his people that the radio was not the making of the devil and that it could in fact be used to broadcast and spread the Quran. In 2003, the Arab region is found to be still lagging in modern technologies adoption (UNDP, 2003). Thus, in a little more than 11 centuries, the Arabs were transformed from leaders to adopters, then to late adopters as far as technologies are concerned.

The Arab world is taken to mean the 22 members of the Arab League, accounting for more than 300 million people with an economy of 700 billion dollars. Although most Arabs practice Islam, they represent less than one third of all Muslims. The Arab world is often thought of as economically prosperous due to its oil resources; yet its total GDP is lower than that of Spain (UNDP, 2003).

Arab countries share language and culture but differ in size, geography, demographics, government type, etc. In terms of spending in and use of IT, statistics portray the Arab world as far from being up-to-date (UNDP, 2003).

This article raises the question of whether existing research models are appropriate to explain this delay. If certain characteristics proper to the Arab culture play an important role in explaining the delay in IT adoption, then we anticipate the answer to be that no existing model is able to adequately study the IT adoption phenomena in the Arab world.

THE ARAB WORLD AND IT

It is a recognized fact that the Arab world lags behind in terms of IT (Aberdeen, 2001; UNDP, 2003). (See Figure 1.) In 2001, Ajeeb reported that the number of Arab Web sites reached 1.24% of the Arab population (Ajeeb, 2001), it predicted that the number of Internet users would equal 5 million by 2001, and reiterated DITnet’s March
2000 prediction that Internet users would number ten to twelve million by the end of 2002. In fact, there were a reported 8.2 million users by the end of 2002.

To date, no Arab country has been able to develop its electronic commerce capabilities to the extent seen in the West.

**REASONS FOR THE LAG**

The lag can partly be explained by the delay with which technologies have traditionally reached Arab countries. Davison et al. (2000) suggested several other reasons: a perceived incompatibility between local cultures and imported technologies, a preference for autonomy and independence with respect to technology, and a lack of economic resources to acquire technology.

The first two of are plausible as is it often the case that IT stumbling blocks occur not because of technical reasons but rather because of human and social obstructions. The third reason can be excluded for the six Gulf countries which claim per capita revenues of nearly five times the average of the rest of the Arab countries. The rate of adoption of the Internet for these countries is up to fifteen times that of the rest of the Arab world.

Other factors also explain the low rate of Internet penetration in Arab nations as compared to the rest of the world. In these nations, the rate of penetration is essentially measured based on only one part of society: men.

**The Arab Woman**

Two thirds of the 65 million illiterate Arabs are women. Women make up only 4% of all Arab Internet users while in Europe women make up 42% on average. The UNDP states that the condition of women is one of the three main factors explaining the current state of economic development in the Arab world.

In that more than 40% of women in Arab countries are denied participation to the technological revolution, Arab nations are failing to integrate a considerable part of their human resources in their future economies.

**Information and IT**

When Arab countries invest in IT, they do so mainly in hardware. While this may be a characteristic of developing countries, it may also be viewed as Arabs’ distrust of anything immaterial. Software on the other hand is associated with innovation, creativity, and the free flow of information and knowledge, qualities that the Arabs have been found lacking (UNDP, 2003). Thus, not only Arabs are increasing their dependence to the West being consumers of hardware, they seem to be passive users of the software and intelligence produced elsewhere.

This issue leads to the tight relationship between information (and not IT, let alone hardware) and democracy and freedom. If Arab countries are truly “information shy” (Henry, 1998), then what information is to be shared and circulated by IT? Therefore, the Arab does not see what use he could make of IT and would therefore not consider it an instrument of development.
Figure 2a. While hardware is projected to represent less than 35% of all IT investments in the world...

Figure 2b. ...it will continue to represent more than 70% of all IT investments in the MENA region (Source: Aberdeen, 2001).

Culture

Culture is “[the] collective programming of the mind that distinguishes the members of one group of people from those of another” (Hofstede, 1991). For Ibn Khaldun, 14th century Arab sociologist, man is son to his habits and his environment, not to his nature and his moods.

Hill et al. (1998) claim that the Arab culture is one of the most complex of the world. It is complex for several reasons. Though the majority of Arabs are Muslim, many are Jewish or Christian while others are agnostic and even atheists. Many actions and symbols assume a particular significance. Often in traditional Arab cultures a handshake or the spoken word are seen as more legally binding than a signed contract and the written word has less significance than in other cultures.

For Arabs bartering is very different from what the West calls negotiating. The numerous steps involved in bartering hold social meanings, the least of which is to learn about and get more acquainted with the other so as to build trust in terms of the product quality. Selling and buying are social acts not replicated by technology.

In his study of Qatari and Emirati cultures, Solberg (2002) found that “wasta” (social connections and string-pulling) could be used to lower the level of uncertainty during business transactions. Even when more direct means are available, Arabs prefer wasta because of the human contact it offers. Wasta is a way of life that builds upon human interactions, a major part of an Arab’s life, which she or he may not be willing to sacrifice to technology.

Arabs tend to be more talkative than Westerners; they are fond of sentences with several adjectives, the number of which, combined with the length of sentences and the tone of the voice, make up an integral part of the message. For example, when a guest is offered a cookie, it is customary to not accept too keenly or quickly. It is up to the host to insist (invoking God) to convert the initial refusal into a yes. Arabs rarely say yes or no, rather, they are inclined to answer using less committed terms such as “usually” and “Inshallah”. The design of computer interfaces with only the “yes”, “no” and “cancel” options is clearly rooted in other cultural norms.

Theories and Research Models

Many research models are available to study IT adoption. Flowers et al. (1975), for example, built
upon the framework of Graves (1970) in which people would evolve through levels of “psychological existence” reflecting their personal values and lifestyles, which vary with their cultural programming.

Research using this framework suggests that dominant value systems of Middle-Eastern managers are conformist, sociocentric and existential and their dominant decision-making style is consultative (Amirshahi, 1997).

Another important stream is represented by Hofstede’s work (1991) which considers culture a national trait. The results of Hofstede’s research is a typology of cultures based on five cultural dimensions: power distance, individualism, masculinity, uncertainty avoidance, and long-term orientation.

In research pertaining to IT, some dimensions have been invoked more than others: individualism (Jarvenpaa et al., 1999), power distance (Simon, 2001) and masculinity (Hofstede, 2001b). Findings from Hofstede (2001b) show a relationship between masculinity and fax machines, and femininity and cellular phones. Like authority, a masculine trait, a fax is not interactive; and like exchange and dialogue, which are feminine traits, a cellular phone is interactive.

Certain Arab values could be mapped onto Hofstede’s cultural dimensions. What one calls “maslaha al amma” (public interest) happens to mean the exact opposite of individualism. In the Arab culture, public interest could even be extended to the notion of Umma (nation). Family commitments and responsibility towards authority figures and political leaders are often held as priorities over other matters. Arabs from the Gulf swear allegiance to the family, clan, tribe, religion, and nation (Abbas, 1987). Usunier (1993) suggests that Arab cultures tend to resemble cultures that are focused on the present and for whom the future (destiny) is inevitably predestined (Mektoub) by God (Inshallah: God willing). For Patai (1973), fatalism is a predominant trait of the Arab culture. If past-oriented cultures are more passive than future-oriented cultures, the Arabs should have an attitude that can be summarized as “if it is written then it will happen” henceforth they will not feel the urge to go toward the technology and will expect it to come to them.

LIMITATIONS OF EXISTING MODELS

The Absence of Arab Dimensions

If in Hofstede’s model, the first four dimensions are applicable to other cultures, it would be useful to more closely study Arab values in order to include them in existing models.

Cultures are likely to diverge rather than come together as we would assume on the basis of globalization and the democratization of media. Hofstede (2001b) has shown that cultures are diverging rather than converging. More specifically, current world events do not lend support to the hypothesis that Arab and Western cultures are closing in.

Current research models would therefore benefit from including Arab dimensions; otherwise they would risk being applicable only to those cultures for which they were developed or for the Arab culture to continue being perceived through the lens of another culture. This could be done following the example of the Far East cultures. These cultures were really only taken into account into Hofstede’s model with the addition of the fifth dimension, Confucian dynamism, thereafter called long-term orientation. It was added in order to complete Hofstede’s model since it was the dimension that distinguished western and far-eastern cultures.

The Implicit Cultural Homogeneity

Differences exist among Arab countries (Solberg, 2002), and even between different regions of a same country. This is even more noticeable as
the borders between Arab countries were often drawn arbitrarily during colonization.

Fundamental cultural differences exist between Sunnis and Shiites, among them masculinity (Noland, 2003). Sunnis are considered more “triumphant” (sic) while Shiites focus more on suffering. The two strains of Islam could therefore be considered at the extremes of the masculinity dimension although scores for this dimension place the Arab world at the midpoint of the scale. However, even Sunnis bear certain feminine traits such as seeking good relationships with leaders and cooperating with colleagues in order to foster equally good relationships with them (Weir, 2001).

Differences can also be found between the cultures of the Gulf and other nations of the Arab world. The discovery of oil in the 1930s resulted in an odd mix of traditional and tribal cultures with modern cultures. This mix did not manifest itself in North Africa resulting in the Arab world composed of countries generally having either people or oil\(^{10}\). Furthermore, North African countries have not always had Arab identity. Unlike the Gulf countries, they became Arabs at the same time as Muslims.

**The Cultural Barrier**

Arabs, Arab culture, and Arab religion have always been a black box for the Westerner. For instance, some westerners have a near-Manichean tendency to dichotomize (Good vs. Evil, Them vs. Us, etc.) and are fond of clash-of-civilizations theories (see Huntington, 1996). El-Solh and Mabro (1994, p. 14) show how the application of over-simplistic dichotomies to the study of gender relations in the Middle East have left out much of the flexibility women might have in actual terms. The simple idea that the sexes have a complementary role in Arab society has often been misunderstood in Western studies as a binary opposition with a tacit hierarchy.

Furthermore, if by fatalism it is understood that humans have no say in their destiny, then the concept of time orientation of Patai (1973) originates from a fundamentally inaccurate understanding of a large part of the Arab world. Islam is founded on the promise of salvation (Jenna) or the threat of damnation (Jehenna). Debates on the “kadha and kadar”\(^{11}\) (destiny and predestination) continue to fuel the controversy that divides theologians on the issue. Suffice it to mention that Arab societies do not necessarily neatly fit within cultures anchored in either the past or the present.

**CONCLUSION**

Significant contradictions may exist between how Arabs perceive themselves and how they are perceived by others. For example, Hill et al. (1998) argue that the adoption of technologies is rarely in quest of imitating the West. But Ibn Khaldun maintains that imitation is characteristic of Arabs. Abbas and Al-Shakti, (1985) even suggested that Arab executives are the product of cultural values that tend to produce more followers than leaders. Yet, imitation is sometimes suspicious in the Muslim’s eye as some believe that imitation of the non-Muslims is “haram” (sinful) While it is also believed that imitating non-Muslims is permissible, the average Arab Muslim sometimes wonders when imitation stops being haram and starts being hallal (allowed)\(^{12}\). How to reconcile these points of view?

Two theories seem promising in that they may complement the research models we reviewed here. Social Identity Theory recognizes that cultural layers exist that describe different levels of programming (social, national, regional, religious, contextual, organizational, etc.).

Abdul-Gader and Kozar (1995) borrowed the construct of technological alienation from psychosociology to explain certain purchase and use decisions of IT. They showed that alienated
individuals resist any kind of technology adoption. More generally, Value-Expectancy Theory (Feather, 1995) promises to enrich the debate on IT adoption by Arabs since it addresses the issue of the value attributed to things by individuals and their expectations, founded or not, such as their resistance to the possible danger of technological and cultural dependence. This is all the more valid that Arabs view IT as a technology, not as a medium of knowledge and of accessing knowledge, something they need direly as evidenced by the conclusions of the UNDP (2003).

REFERENCES


**KEY TERMS**

**Arab World:** The Arab world is taken to include all 22 countries members of the Arab League.

**Culture:** According to Hofstede (1991), it is “[the] collective programming of the mind that distinguishes the members of one group of people from those of another”. For the 14th century Arab scholar, Ibn Khaldun, man is son to his habits and his environment, not to his nature and his moods. In all the literature about culture, there is a common understanding that culture is an abstraction from concrete behaviour but is not behaviour itself. Hofstede’s typology includes five cultural dimensions:

- **Individualism:** as opposed to collectivism, describes the degree to which a culture relies on and has allegiance to the self or the group.
- **Power Distance:** reflects the degree to which a culture believes how institutional and organizational power should be distributed.
- **Masculinity:** indicates the degree to which a culture values such behaviours as assertiveness, achievement, acquisition of wealth, social supports and the quality of life.
- **Uncertainty Avoidance:** refers to the extent to which a culture feels threatened by
ambiguous, uncertain situations and tries to avoid them.

- Long-Term Orientation: reflects the degree to which a culture is oriented toward the search for virtuous behaviour rather than being oriented towards the search for truth.

Information Technology: Hardware, software, network and services related to the use and operation of equipment with the aim of processing and communication of analogue and digital data, information, and knowledge. These include computers and computer applications such as the Internet, Intranets, Extranets, Electronic Data Interchange, electronic commerce, mobile and fixed lines, etc.

ENDNOTES

1 Most technological innovations are adopted by the Arabs with some delay. For example, if the printing press was invented around 1450, the first Arab book to have been printed was the Quran, in 1537 in Venice. Arabic character-based printing reached the Arabs only in 1728 under the reign of Ahmed III (1703-1730) when the Arabs were occupied by the Ottoman Empire. Bukhara (today’s Uzbekistan) delayed the introduction of printing until 1917 (See Khalid, 1994). It is as though in the Muslim world, the advent of printing broke the ulama’s long-established control of knowledge. Similarly to the Bible and the printing press, no longer could knowledge be obtained through someone “more knowledgeable”. It could be read and understood - or misunderstood—in libraries, schools, and even homes. Later, while e-mail was invented in 1971, the first Arabic Web-based mail solution was released in December 1998 (see http://web1.maktoob.com/maktoob/press1998/press1998-1.html). See more on Kaitlin Duck Sherwood’s site, http://www.webfoot.com/advice/WrittenArabic.html.


4 For Muslim Arabs, this may be explained historically and religiously by the fact that when the Divine message was delivered, the Angel Gabriel dictated the Quranic verses to the Prophet Muhammad. In all pre-modern times, documents were not copied; they were memorized, where there was no other way to preserve them.

In his book, “The Saddam Years” (Fayard, 2003), Saman Abdul Majid, personal interpreter to the deposed dictator, explains how, in 1993, President Clinton sent a secret agent to Iraq to suggest that a new leaf be turned over and that discussions be resumed. Saddam did not immediately answer, an act that Clinton took as a refusal. That file was then closed. In fact, Saddam was expecting a more solid and thought-out proposition to be put forward, and was surprised that Clinton did not come through with one. This miscommunication between two men of very different cultures has had the now all-too known consequences.

6 It is assumed that most readers are familiar with Hofstede’s work. Due to space limitations, details of his work will not be elaborated here. For more information, the reader is referred to Hofstede (2001b).

7 Also of interest is the GLOBE (Global Leader and Organizational Behavior Effectiveness) project which seeks to determine the relationship between leadership and societal culture (House et al., 2002). GLOBE uses
nine dimensions, six of which are the same as Hofstede’s: uncertainty avoidance, power distance, institutional collectivism, in-group collectivism, gender egalitarianism, assertiveness, future orientation, performance orientation, and humane orientation. Using these dimensions, Gupta et al. (2002) identified 10 “cultural clusters”: Asia, Anglo, Arab, Germanic Europe, Latin Europe, Eastern Europe, Confucian Asia, Latin America, Sub-Saharan Africa and Nordic Europe. Unfortunately, often in GLOBE papers, the Arab cluster is composed of very few countries (see Abdalla and Al-Homoud, 2001). Other countries, such as Turkey, have disconcertingly been found in such “Arab clusters” (see Kabasakal & Bodur, 2002) ignoring the fact that Turkey is not a member of the Arab League and overlooking its claim to European Union membership. Perhaps even in this order, especially if one takes into account the famous proverb in which an Arab states: “Me, against my brother; my brother and I against our cousin; my brother, my cousin and I against all others.” Was it not written of Ishmael (and by extension of all Arab children) in the Bible that “And he will be a wild man; his hand will be against every man, and every man’s hand against him; and he shall dwell in the presence of all his brethren” (Genesis 16:12). This, however, does not seem to hold any longer given the events that took place between Iraq and Kuwait in 1990-1991 as well as since then. Judging by the lack of cultural sensitivity of today’s’ graphical user interfaces and icons, that technology is all but coming to the Arab culture. The only exception being Iraq. This highly philosophical issue revolves around the overly simplified question that if God had already decided on much of what we humans will do on Earth even before we were created, what is the Last Judgment about?

One has to remember that the digging of the trenches around Madina in the battle of Al-Ahzab, which was led by the prophet Mohamed (Peace Be Upon Him) and his companions as a defense system, and the application of Omar Ibn Al-Khattab, the second khalif after the prophet, of the filing system in the Islamic State, for organizing taxes on cultivated land and payroll of the soldiers are examples of imitation as they were borrowed from Non-Muslim Persians (see more on this subject on www.islam1.org/khutub/Imitation_of_None-Muslims.htm, accessed April 13, 2004).
Chapter 1.11

Information Technology Acceptance across Cultures

Amel Ben Zakour
FSJEGJ, Tunisia

ABSTRACT

This chapter introduces national culture as a possible factor accounting for the differences in information technology adoption and use between countries. Based upon culture theory and the technology acceptance model (TAM), the author offers a conceptual model aiming at better understanding IT acceptance across countries of different cultures. It has been argued that six value dimensions—individualism/collectivism, power distance, uncertainty avoidance, masculinity/femininity, high/low context, and polychronism/monochronism—act as moderators of the TAM relationships. Furthermore, the author aims at helping IT designers and IT managers all over the world to understand why certain national cultural values may be congruent or not with the IT to be designed or implemented.

INTRODUCTION

With the globalization context, companies all over the world (developed and less developed ones) are using information technologies (IT) that have become more and more sophisticated and complex. These technologies provide huge opportunities to gain a competitive advantage since information could be obtained, processed and transmitted at very low costs (Porter & Millar, 1985). Nevertheless, in order for companies to take full advantage of IT, they have to understand the new challenge provided by these IT (Tapscott & Caston, 1994). Actually, even though organizations adopt IT that best fit their business activities, they cannot guarantee performance leveraging unless the organization members appropriately use it. According to Agarwal (1999), “Acquiring appropriate IT is necessary but not sufficient condition for utilizing it effectively” (p. 85). With the globalization incurring unlimited interconnection possibilities and an increasing number of partnerships, firms belonging to less-developed countries are investing massively in new infor-
Information Technology Acceptance across Cultures

Information technologies that are expected to improve their competitiveness. Nevertheless, we notice that although these technologies are voluntarily purchased by organizations, they are not fully used or accepted in the work context at the individual level. Differences in IT usage and adoption are reported in many descriptive studies pertaining to IT implementation in different countries and contexts (Danowitz, Nassef, & Goodman, 1995; Goodman & Green, 1992). The main causes of these differences that have been identified are technical, economic, or managerial.

Our main concern in this chapter is to shed light on an issue that deserves a closer study which is the cultural factor. Indeed, the latter is expected to be an interesting explanation for the differences in IT adoption and use between countries. Furthermore, culture has been used as an explanatory factor of several managerial and organizational issues (Fisher & Smith, 2003; Thomas & Pekerti, 2003; Laurent, 1983; Hofstede, 1985; Silvester, 1990; Hernandez, 2000). The literature in the information systems (IS) field provides few studies attempting to explore the nature of the relationship between culture and IT implementation. Furthermore, most of the prior cross-cultural studies on IS hardly have focused on the individual behavior toward IT; they generally have focused on IT transfer (Hill, Loch, Straub, & El-Sheshai, 1998), on organizational characteristics related to IT implementation (Robey & Rodriguez-Diaz, 1989), or on end-user computing characteristics (Igbaria & Zviran, 1996). Moreover, these studies have focused on the concept of culture at the macro level of analysis (i.e., they attempted to compare IS-related behaviors in different countries, supposing that each country is characterized by a different set of cultural values). For example, when studying IT diffusion in Japan and the USA, Straub (1994), using Hofstede’s cultural classification of 51 countries and three regions, supposed that the Japanese are high in uncertainty avoidance (ranked 7th) and the Americans are low in this dimension (ranked 43rd). Even though there is accumulated evidence pointing to national culture as a factor influencing IT adoption and acceptance, few studies have handled this issue at the individual level. Therefore, the main objective of this chapter is to provide a conceptual framework that examines the influence of national culture, though a macro-level construct, on IT adoption by integrating specific cultural value dimensions to technology acceptance model (TAM) (Davis, Bagozzi, & Warshaw, 1989). This framework is aimed at being used at the individual level since cultural dimensions characterizing national culture (i.e., individualism/collectivism, power distance, masculinity/femininity, uncertainty avoidance, high/low context of communication and polychronism/monochronism) are value dimensions considered as individual psychological dispositions and TAM is designed to capture acceptance at the individual level.

We have divided the chapter into four parts. The first part is devoted to a review of comparative and international studies in the IS field. The second part presents an overview on culture theory, which is found in the literature of various disciplines such as anthropology or cross-cultural psychology. The third part focuses on the presentation of TAM (Davis et al., 1989), judged to be the most parsimonious model of information technology usage, as well as TAM-relevant extensions. The fourth part is aimed at presenting the rationale sustaining the conceptual framework developed to explain how national culture could influence IT usage.

STUDIES INVOLVING NATIONAL CULTURE AND INFORMATION SYSTEMS

Prior researches focusing on cultural issues in the IS field were conducted mainly through two main and different perspectives. According to Ronen and Shenkar (1988), the *emic* perspective “studies
behavior from within a single culture-system” (p. 72), whereas the etic perspective focuses on universals and studies behavior “from a position outside the system and in comparison with various cultures” (p. 72). Early and Singh (1995) have provided more comprehensive classifications of international and intercultural studies. Early and Singh (1995) have proposed four research approaches to study these issues; namely, unitary form, gestalt form, reduced form, and hybrid form. The difference between these approaches stems from “examining whole systems rather than their components” (Early & Singh, 1995, p. 330).

The unitary form emphasizes the understanding of a particular cultural group using the specific constructs and concepts of this group. Actually, this kind of approach is not so much interested in cultural dimensions per se as much as it focuses on knowing how a certain behavior is manifested in a particular setting. There is no attempt to establish universals based upon such a perspective.

The gestalt form, unlike the unitary system form, focuses on the comparison of the phenomenon with a similar one in different contexts, and “constructs and relationships are derived from general principles rather than from the systems themselves” (Early & Singh, 1995, p. 330). This approach is appropriate when the researcher is concerned with the generalizability of a model across nations.

The reduced form examines the phenomenon of interest with regard to a specific feature of culture, such as the case dealing with one cultural dimension when hypothesizing about a certain behavior. Generally, comparative studies that explain the differences in the applicability of a model by the country of origin actually are using the reduced form. The cultural explanation is provided ad hoc.

The hybrid form goes beyond comparison in search of similarities and differences, and it allows researchers to explain how certain aspects of a culture can explain the phenomenon. “The reliance on nation or culture as a ‘black box’ is abandoned in favor of a more precise specification of theoretical relationships” (Early & Singh, 1995, p. 334).

In order to have a better visibility of cross-cultural researches in the IS field, we present in Table 1 a summary of cross-cultural studies on IS, grouped according to Early and Singh’s classification.

Our current research follows more the hybrid form since we integrate to TAM a set of universal cultural dimensions deemed to characterize national culture.

### Table 1. Overview of comparative and national studies on IS

<table>
<thead>
<tr>
<th>Authors</th>
<th>IS Research Area</th>
<th>Countries</th>
<th>Key Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Igbaria and Zviran (1996)</td>
<td>End-user computing (EUC)</td>
<td>USA Israel Taiwan</td>
<td>EUC characteristics (individual and organizational), perceived usefulness, effectiveness of EUC, satisfaction and acceptance vary across countries.</td>
</tr>
<tr>
<td>Straub, Keil, and Brenner (1997)</td>
<td>IT individual acceptance</td>
<td>USA Japan Switzerland</td>
<td>TAM explains the use of e-mail in the USA and Switzerland but not in Japan. The cultural Japanese orientation may limit e-mail use and dissociate usefulness from use.</td>
</tr>
</tbody>
</table>

*continued on following page*
### Information Technology Acceptance across Cultures

**Table 1. Continued**

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Context &amp; Methods</th>
<th>Countries/Regions</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Leidner and Carlsson (1998)</strong></td>
<td>Executive Support System (ESS) adoption</td>
<td>Sweden, USA, Mexico</td>
<td>National culture influences the perception of the benefits of ESS on decision-making processes.</td>
</tr>
<tr>
<td><strong>Hill et al. (1998)</strong></td>
<td>IT transfer</td>
<td>Jordan, Egypt, Saudi Arabia, Lebanon, Sudan</td>
<td>Cultural factors identified in the Arab world such as family and kinship obligations, communal world view, religion, valuing the past or the concept of time, are powerful for predicting the outcomes of technology transfer.</td>
</tr>
<tr>
<td><strong>Unitary Form Studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martinsons and Westwood (1997)</td>
<td>Management Information System (MIS) adoption and use</td>
<td>China</td>
<td>The cultural Chinese dimensions of paternalism, personalism, and high context communications shape the use of MIS in China.</td>
</tr>
<tr>
<td>Gamble and Gibson (1999)</td>
<td>Decision making and use of information</td>
<td>Hong-Kong</td>
<td>The collectivism characterizing the Chinese (paternalistic and autocratic) leads to the distortion or restriction of the information produced by the financial information system.</td>
</tr>
<tr>
<td>Rowe and Struck (1999)</td>
<td>Media choice</td>
<td>France</td>
<td>The use of asynchronous media is more associated with a cultural value oriented toward reactivity. The media that provide more redundancy in the feedback are more associated with relation than with task-oriented values.</td>
</tr>
<tr>
<td><strong>Reduced Form Studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Straub (1994)</td>
<td>IT diffusion</td>
<td>Japan, USA</td>
<td>Since the Japanese are characterized by a high level of uncertainty avoidance, they view e-mail as more information-rich and socially present than the Americans. The structural features of the Japanese language make e-mail difficult to use.</td>
</tr>
<tr>
<td>Rice, D’Ambra, and More (1998)</td>
<td>Media choice and assessment</td>
<td>Hong-Kong, Singapore, Australia, USA</td>
<td>The collectivists prefer face-to-face interaction more than the individualists do. The individualists consider telephone and business memo richer than the collectivists do. There is no difference in the perception of equivocal situations between individualistic people and collectivistic ones.</td>
</tr>
<tr>
<td>Hasan and Ditsa (1999)</td>
<td>IT adoption</td>
<td>West-Africa, The Middle-East Australia</td>
<td>Differences were reported in IT adoption along eight cultural dimensions, which are Hofstede’s four cultural dimensions, polychronism/monochronism, time orientation, context, and polymorphic/monomorphic.</td>
</tr>
<tr>
<td>Hofstede (2000)</td>
<td>IT adoption</td>
<td>56 countries</td>
<td>The higher a country scores on uncertainty avoidance, the slower it will be in adopting a new technology. IT adoption also is influenced by masculinity/femininity and individualism/collectivism dimensions.</td>
</tr>
<tr>
<td><strong>Hybrid Form Studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Srite and Karahanna (2006)</td>
<td>IT adoption</td>
<td>30 countries</td>
<td>Hofstede’s four cultural dimensions have moderating effects on TAM causal relationships.</td>
</tr>
</tbody>
</table>
CULTURE THEORY

Even though originally rooted in anthropology, a population-level discipline, culture has been defined and researched by many other disciplines such as cross-cultural psychology. Culture has been defined according to several perspectives. Definitions go from the most complex and the most comprehensive (Kluckhohn, 1962) to the simplest (Hofstede, 1997; Triandis, 1972). According to Kluckhohn (1962), “Culture consists of patterns, explicit and implicit, of and for behavior acquired and transmitted by symbols, constituting the distinctive achievement of human groups, including their embodiments in artifacts; the essential core of culture consists of traditional (i.e., historically derived and selected) ideas and especially their attached values; culture systems may, in one hand, be considered as product of action, on the other hand, as conditioning influences upon further action” (p. 73). Hofstede (1997) defines culture as “the collective programming of the mind which distinguishes the members of one group or category of people from another” (p. 5).

Researchers in comparative and intercultural management most of the time use the concept of organizational culture or national culture. Nevertheless, they omit the fact that individual behaviors and attitudes in an organizational context could be influenced by other kinds of cultures. Indeed, culture is a multi-level phenomenon that could be approached according to different levels such as region, ethnic group, religion, language, nation, profession, firm, gender, social class (Hofstede, 1997; Karahanna, Evaristo, & Srite, 2005; Schneider & Barsoux, 2003). Culture also could be defined according to continental or political belonging (Lévi-Strauss, 1985). Furthermore, these different cultures interact with each other. For example, several ethnic groups are found in India.

In the present study, we are interested in national culture since it has been shown to influence management and organizations (Hernandez, 2000; Hofstede, 1985; Laurent, 1983; Silvester, 1990; Zghal, 1994). In a comparative study involving the USA, Indonesia, Japan, and nine countries from Western Europe, Laurent (1983) has found that managers from these countries have different behavioral patterns. For example, the Americans believe that the main goal of a hierarchical structure is to organize tasks and facilitate problem resolution, whereas for the Japanese, the Indonesians, or the Italians, the goal is to highlight the authority structure. In the same vein, Hofstede (1985) has demonstrated that the employees’ perceptions of the organization and its management depend on the intensity of two national values: power distance and uncertainty avoidance. Four implicit patterns of organization conception are defined: (1) the pyramid pattern (both power distance and uncertainty avoidance) are high, which corresponds to the hierarchical bureaucracy, quite common in France and the Mediterranean countries; (2) the well-oiled machine pattern (low power distance and high uncertainty avoidance), which corresponds to an impersonal bureaucracy found, for example, in Germany; (3) the village market pattern (both power distance and uncertainty avoidance are low), which is found in Great Britain and Nordic countries; and (4) the family pattern (high power distance and low uncertainty avoidance), which is found for example in African and Asian countries. Hernandez (2000) also has shown that African management systems are influenced by a national cultural trait: paternalism. The latter corresponds to high power distance and relationships based upon mutual trust. Paternalism also has been found to be perceived positively in Tunisian culture contrarily to American culture, which is focused on contractual relationships or Dutch culture, which is focused on consensual relationships (D’Iribarne, 1989). In a comparative study between France and Germany, Silvester (1990) has confirmed the results obtained by Hofstede (1985), showing that the French organization could be assimilated to a pyramid and the German one.
Information Technology Acceptance across Cultures

to a well-oiled machine. For example, they found that French managers have a prescriptive kind of work, whereas German managers have cooperative kind of work through which employees could learn and build a collective identity.

Several sets of dimensions have been developed to characterize the concept of national culture. Table 2 provides an overview of the most known cultural dimensions found in several fields of studies.

The analysis of the culture frameworks in Table 2 reveals an overlap of cultural dimensions identified by the various authors. Even though the definitions of these dimensions are not fully convergent in all the cases, still they show some similarities. Table 3 gives an overview of the similarities that exist between these dimensions. The current research will use the following dimensions: individualism/collectivism, masculinity/femininity, uncertainty avoidance, power distance (Hofstede, 1997), high context/low context (Hall, 1989), and time perception (Hall, 1989; Trompenaars & Hampden-Turner, 1998).

Several reasons lead us to investigate national culture through these six dimensions.

First, these dimensions rely on variables that are linked more directly to social and organizational process: they focus on human values rather

---

Table 2. Overview of the most known cultural dimensions

<table>
<thead>
<tr>
<th>Cultural Dimensions</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Distance</td>
<td>Hofstede (1997)</td>
</tr>
<tr>
<td>Individualism/Collectivism</td>
<td>Hofstede (1997)</td>
</tr>
<tr>
<td>Masculinity/Femininity</td>
<td>Hofstede (1997)</td>
</tr>
<tr>
<td>Uncertainty Avoidance</td>
<td>Hofstede (1997)</td>
</tr>
<tr>
<td>Long-term Orientation</td>
<td>Hofstede (1997)</td>
</tr>
<tr>
<td>Confucian Work Dynamism</td>
<td>Chinese Culture Connection (1987)</td>
</tr>
<tr>
<td>Conservatism</td>
<td>Schwartz (1994)</td>
</tr>
<tr>
<td>Intellectual autonomy</td>
<td>Schwartz (1994)</td>
</tr>
<tr>
<td>Affective autonomy</td>
<td>Schwartz (1994)</td>
</tr>
<tr>
<td>Hierarchy</td>
<td>Schwartz (1994)</td>
</tr>
<tr>
<td>Egalitarianism</td>
<td>Schwartz (1994)</td>
</tr>
<tr>
<td>Mastery</td>
<td>Schwartz (1994)</td>
</tr>
<tr>
<td>Harmony</td>
<td>Schwartz (1994)</td>
</tr>
<tr>
<td>Universalism/Particularism</td>
<td>Trompenaars and Hampden-Turner (1998)</td>
</tr>
<tr>
<td>Individualism/Communitarianism</td>
<td>Trompenaars and Hampden-Turner (1998)</td>
</tr>
<tr>
<td>Neutral/Emotional</td>
<td>Trompenaars and Hampden-Turner (1998)</td>
</tr>
<tr>
<td>Specific/Diffuse</td>
<td>Trompenaars and Hampden-Turner (1998)</td>
</tr>
<tr>
<td>Achievement/Ascription</td>
<td>Trompenaars and Hampden-Turner (1998)</td>
</tr>
<tr>
<td>Attitudes to time</td>
<td>Trompenaars and Hampden-Turner (1998)</td>
</tr>
<tr>
<td>Attitudes to environment</td>
<td>Trompenaars and Hampden-Turner (1998)</td>
</tr>
<tr>
<td>Communication context</td>
<td>Hall (1989); Hall and Hall (1987)</td>
</tr>
<tr>
<td>Perception of space</td>
<td>Hall &amp; Hall (1987); Hall (1989)</td>
</tr>
<tr>
<td>Monochronic/polychronic time</td>
<td>Hall (1989)</td>
</tr>
</tbody>
</table>
than on general beliefs about the way we see the world. Indeed, “culture is primarily a manifestation of core values” (Straub, Loch, Evaristo, Karahanna, & Srite, 2002); therefore, in order to better capture the relationship between culture and behavior, Triandis suggests using values (Triandis, 1972). Second, the first four dimensions (Hofstede’s cultural dimensions) constitute the most used and most recognized dimensions as a whole or separately in studying cross-cultural issues in management and organizations. Because of its global coverage in terms of respondents, it seems that Hofstede’s study has been unrivaled (Smith & Bond, 1999). In fact, the identification of the cultural dimensions was based upon a field study covering a sample of 40 countries in which more than 116,000 questionnaires were collected. Hofstede’s work also has been validated directly or indirectly by many other researchers in different settings (The Chinese Culture Connection, 1987; Hofstede & Bond, 1984; Shackleton & Ali, 1990). The studies that have been conducted

---

Table 2. Continued

<table>
<thead>
<tr>
<th>Nature of people</th>
<th>Kluckhohn and Strodtbeck (1961)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person’s relationship to nature</td>
<td>Kluckhohn and Strodtbeck (1961)</td>
</tr>
<tr>
<td>Person’s relationship to other people</td>
<td>Kluckhohn and Strodtbeck (1961)</td>
</tr>
<tr>
<td>Primary mode of activity</td>
<td>Kluckhohn and Strodtbeck (1961)</td>
</tr>
<tr>
<td>Conception of space</td>
<td>Kluckhohn and Strodtbeck (1961)</td>
</tr>
<tr>
<td>Person’s temporal orientation</td>
<td>Kluckhohn and Strodtbeck (1961)</td>
</tr>
</tbody>
</table>

---

Table 3. Similarities in cultural dimensions

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Power distance</td>
<td>Hierarchy / Egalitarianism</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Individualism/ Collectivism</td>
<td>Autonomy</td>
<td>Individualism/ Communitarianism</td>
<td>-</td>
<td>-</td>
<td>Relational Orientation</td>
</tr>
<tr>
<td>Masculinity/ Femininity</td>
<td>Mastery/ Harmony</td>
<td>Achievement/ Ascription Inner-directed/ Outer-directed</td>
<td>-</td>
<td>-</td>
<td>Man-Nature Orientation</td>
</tr>
<tr>
<td>Uncertainty Avoidance</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Long-term Orientation</td>
<td>Conservatism</td>
<td>Attitudes to Time</td>
<td>Confucian Work Dynamism</td>
<td>Time Perception</td>
<td>Time Orientation</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>Specific/ Diffuse</td>
<td>-</td>
<td>Space (personal space and territory)</td>
<td>Space Orientation (public, private)</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>High /Low Context</td>
<td>-</td>
</tr>
</tbody>
</table>
Information Technology Acceptance across Cultures

since Hofstede’s work (The Chinese Culture Connection, 1987; Schwartz, 1994; Trompenaars & Hampden-Turner, 1998) exploring the national culture through values have sustained and amplified his findings rather than having contradicted them (Smith & Bond, 1999). The two dimensions related to time orientation and communication context are based upon Hall’s well-established studies on intercultural communications. Indeed, cross-cultural studies of styles of communication, briefly reviewed by Smith and Bond (1999), reveal a divergence between societies in several aspects of communication and provide evidence sustaining Hall’s contention about high/low context (Gudykunst & Ting-Toomey, 1988).

In the following paragraphs, we are going to give more explanations about the origins of these cultural dimensions.

**Power Distance (PDI)**

According to Hofstede (1997), individuals in high PDI societies accept the unequal distribution of power, which is what Schwartz (1994) calls hierarchy. The latter is linked to the inequality of the distribution of power, roles, and resources. Conversely, in low PDI societies, individuals do not accept this inequality. Indeed, in an organizational context, subordinates can criticize their superiors, and superiors are supposed to take into account this criticism, since everyone is free to express his or her disagreement about any issue.

In a low PDI organization, the hierarchical system depicts an inequality of roles that may change over time (Hofstede, 1997). This kind of behavior is not tolerable in a society that emphasizes authority. What Schwartz identifies as egalitarianism (also egalitarian commitment) (Schwartz, 1994) can be assimilated to the concept of low PDI advocated by Hofstede.

Egalitarianism dimension is driven by values expressing the transcendence of selfish interests and leading to a voluntary commitment to promote welfare of other individuals (e.g., social justice or loyalty). Even though Schwartz explains the opposition hierarchy/egalitarianism by psychological concepts (i.e., self-enhancement/self-transcendence), Hofstede explains PDI dimension, relying more on social contexts in which individuals have learned to accept or reject the inequality of relationships between people (such as country history, ideologies, language). These two dimensions are fairly comparable and valid to characterize national culture.

**Individualism/Collectivism (I/C)**

Individualism concept takes its roots in the Western world. Indeed, in Britain, the ideas of Hobbes and Adam Smith about the primacy of the self-interested individual sustain this concept. In contrast, Confucianism in the Eastern world, emphasizing virtue, loyalty, reciprocity in human relations, righteousness, and filial piety, underlies a collectivistic view of the world. Even though some societies are characterized as individualistic and others as collectivistic, they have to deal with both individualistic and collectivistic orientations. These orientations co-exist, and what makes the difference among societies is the extent to which they emphasize individualistic and collectivistic values.

Individualism/collectivism has been criticized by several disciplines and especially psychology (for a review, see Kagitçibasi, 1997). Nevertheless, it is still the focal dimension in cross-cultural studies and has been used most often as an explanatory variable (Schwartz, 1994).

Hui and Triandis (1986) define collectivism as “a cluster of attitudes, beliefs, and behaviors toward a wide variety of people” (p. 240). Seven aspects of collectivism has shown to be relevant in characterizing it, which is the consideration of the implications of our decisions for other people, the sharing of material resources, the sharing of nonmaterial resources (e.g., affection or fun), the susceptibility to social influence, self-presentation and facework, the sharing of outcomes and
the feeling of involvement in others’ lives (Hui & Triandis, 1986).

All the dimensions pertaining to this concept of I/C (Hofstede, 1980), developed by Schwartz (autonomy/conservatism), Trompenaars and Hampden-Turner (individualism/communitarianism) and Kluckhohn and Strodtbeck (relational orientation), deal with one unique issue: the person’s relationship to other people. Schwartz defines autonomy/conservatism according to the person’s embeddedness in the group; meanwhile, Hofstede, Trompenaars, and Hampden-Turner define I/C, emphasizing individual vs. group interest.

Two underlying ideas support the autonomy/conservatism dimension. The first idea explains the autonomy pole: an individual is considered autonomous when he or she is holding independent rights and desires and the relationship he or she has with the others is dictated by self-interest and negotiated agreements. The second idea explains the conservatism pole: an individual is part of a group and his or her significance pertains to his or her participation in and identification with the group (Schwartz, 1994). At the individual level, Schwartz argues that there are two kinds of values: individualistic values, which are hedonism, achievement, self-direction, social power, and stimulation; and collectivistic values, which are pro-social, restrictive conformity, security, and tradition.

According to Hofstede, in individualist cultures, “the interests of the individuals prevail over the interests of the group” (Hofstede, 1997, p. 50). People are supposed to decide and achieve alone. In collectivist cultures, people prefer to decide and act in a group so that responsibility and reward will be shared. “Collectivism is associated with childrearing patterns that emphasize conformity, obedience and reliability” (Triandis, 1989, p. 16).

Individualism/Collectivism construct is considered to be the most important aspect of cultural differences because of its parsimony and its potential to explain variations in economic development (Kagitçibasi, 1997). Indeed, according to Hofstede and Bond (1988), there is a high correlation between individualism and the level of economic growth at the national level.

**Masculinity/Femininity (MAS)**

The MAS dimension was derived empirically from Hofstede’s work on work-related values across countries. Hofstede distinguishes between cultures according to their emphasis on achievement or on interpersonal harmony. In labeling this dimension according to gender, Hofstede refers to the social and culturally determined roles associated with men vs. women and not to the biological distinction. Several other studies have identified almost the same dimension labeled as mastery/harmony (Schwartz, 1994) or achievement/ascription (Trompenaars & Hampden-Turner, 1998). All the dimensions show an obvious overlapping, since they are driven by almost the same set of opposing values. Empirically, a cross-cultural study conducted by Schwartz (1994) shows a positive correlation between MAS and mastery. In one pole, the underlying values are assertiveness, material success, control or mastery, and competition. This is the pole of masculinity (mastery, achievement). In the pole of femininity, the dominant values are modesty, caring for others, warm relationships, solidarity, and the quality of work life.

**Uncertainty Avoidance**

This dimension has been inferred from Hofstede’s survey pertaining to the theme of work stress when he addressed his questionnaire to IBM employees. According to Hofstede, the extent to which individuals tend to avoid uncertainty can differentiate among countries. Indeed, the feeling of uncertainty is something that could be acquired and learned in the diverse institutions of a society such as family, school, or state. Each society will have its proper behavioral model toward this
feeling of uncertainty. Hofstede argues that uncertainty is strongly linked to anxiety. The latter could be overcome through technology, laws, and religion (Hofstede, 1997). In strong uncertainty avoidance societies like South America, people have rule-oriented behaviors. On the contrary, in societies where uncertainty avoidance is weak, people do not need formal rules to adopt a specific behavior. In the same vein, Triandis (1989) makes the difference between loose and tight cultures. He sustains that loose cultures encourage freedom and deviation from norms, whereas in tight cultures, norms are promoted, and deviation from those norms is punished. Shuper, Sorrentino, Otsubo, Hodson, and Walker (2004) also have confirmed that countries do differ in uncertainty orientation. Indeed, they have found that Canada is an uncertainty-oriented society that copes with uncertainty by attaining clarity and finding out new information about the self and the environment, and Japan is a certainty-oriented society that copes with uncertainty by maintaining clarity and adhering to what is already known (Shuper et al., 2004).

**Time Orientation**

The most overlapping dimensions are attitudes to time (Trompenaars & Hampden-Turner, 1998), time perception (Hall, 1989), and time orientation (Kluckhohn & Strodtbeck, 1961), which highlight two main dimensions describing the concept of time: the structure of time (discrete vs. continuous) and the horizon of time reference (reference to the past, present, or future). According to Hall and Hall (1987), a monochronic person runs one activity at a time and associates to each activity a precise time, while a polychronic person can perform many activities simultaneously without preparing an exact schedule, or if it exists, it can be ignored. Studies conducted in this sense (Schramm-Nielsen, 2000) have shown that polychronic time is very specific to the Mediterranean countries, the Middle East, the Arab world, and Latin America. Indeed, in these countries, people assume that time is elastic, and thus, the appointment time and planning are not always respected. Actually, this behavior is considered very normal, because individuals emphasize interpersonal relationships. On the contrary, North-European and Anglo-Saxon countries usually adopt the monochronic time. In these countries, time is divided into segments and rigorously scheduled so that each activity will be assigned a specific and unique time within which it will be performed. Priorities also are classified according to time. Consequently, planning and schedules are sacred and cannot be changed. Trompenaars and Hampden-Turner (1998) called these two types of time systems the sequential and the synchronic. “Time can be legitimately conceived of as a line of sequential events passing us at regular intervals. It can also be conceived as cyclical and repetitive” (Trompenaars & Hampden-Turner, 1998, p. 126).

The time assumption underlying the remaining dimensions (long-term orientation, Confucian work dynamism, and conservatism) pertains to the temporal horizons to which persons refer in their activities: past, present, and future. Confucian work dynamism is based upon values derived from Confucian teachings. In one pole, called by Hofstede long-term orientation, are found values such as perseverance and thriftiness, and in the short-term orientation pole are found values such as personal steadiness and stability and respect for tradition.

Schwartz identifies a value called conservatism that is similar in many aspects to the Confucian work dynamism, since it emphasizes the cohesiveness of the group and the respect for the traditional order of the group. This emphasis is depicted in at least three values characterizing Confucian work dynamism: ordering relationships, personal steadiness, and respect for tradition.
High Context/Low Context of Communication

The nature of the communication context has been highlighted only by Hall (1989) as a relevant dimension capable of discriminating between cultures. In order to better understand the concept of high context/low context of communication, it is crucial to define context and to shed light on the relationship between meaning, context, and communication. According to Hall and Hall (1987), “context is the information that surrounds an event and inextricably bound up with the meaning of that event” (p. 7). A distinction can be made between internal and external contexting. The former depends on past experience internalized into a person or the structure of the nervous system or both (Hall, 1989). The latter is based on the situation and/or setting pertaining to a particular event. Meaning is the interpretation attributed to a particular message or communication. “Words and sentences have different meanings depending on the context in which they are embedded” (Hall, 1983, p. 56). According to Hall (1989), meaning is made up of the message and the two kinds of context, which are the background and preprogrammed responses of the recipient on the one hand and the situation on the other hand. Since information, context, and meaning are the key components in defining high and low context cultures, Hall contends that “the more information that is shared, … the higher the context” (Hall, 1983, p. 56). Therefore, high-context communications are those that rely more on context (shared or stored information) than on transmitted information. In this case, little of the meaning is conveyed in the transmitted part of the message. In high-context cultures, “simple messages with deep meaning flow freely” (Hall, 1989, p. 39). People do not need to elaborate their speech codes because of the largely shared assumptions about the rules of communication. Indeed, the emphasis is more on the manner of transmitting a message than on the message per se. For example, gestures and voice intonation can make the difference between messages of the same content. On the contrary, low-context communications rely more on the transmitted part of the message and less on the context. Americans, Germans, Swiss, and Scandinavians are found to be low-context. In such cultures, characterized by Hall (1989) as fragmented cultures, people have more elaborate codes in communication because of their lack of shared assumptions about communication rules. All the meaning they try to transmit is reflected in the information contained in the literal message. In this case, it seems obvious that there is no need to use different kinds of signals as a guide to interpretation.

TECHNOLOGY ACCEPTANCE MODEL

The technology acceptance model (TAM) is considered a tool designed to understand and measure the IT individual determinants of use. The most comprehensive TAM is designed by Davis et al. (1989). It is based on two key concepts: perceived usefulness (PU), defined as “the prospective user’s subjective probability that using a specific application system will increase his or her job performance within an organizational context” (p. 985); and perceived ease of use (PEOU), defined as “the degree to which the prospective user expects the target system to be free of effort” (p. 985). According to this model, an individual who intends to use a system will have two types of beliefs (PU and PEOU), which influence behavioral intention through attitudes. In addition, PU will have a direct effect on behavioral intention. Finally, external variables will have a direct effect on the two beliefs.

In order to study individual adoption of IT, the present research will be using the TAM. The use of the latter is motivated by theoretical, empirical, and practical considerations. From a theoretical point of view, TAM takes its roots
from the theory of reasoned action (TRA) (Ajzen & Fishbein, 1980). The latter provides a simple model to understand and predict a person’s behavior. It goes from the principle that an individual behaves most of the time in a rational manner, takes into account available information, and assesses consequences of his or her actions implicitly or explicitly (Ajzen, 1988). According to TRA, individual behavior is influenced by behavioral intentions. These intentions at the same time are determined by attitudes toward this behavior and subjective norms. In turn, attitudes and subjective norms are dependent on behavioral and normative beliefs, respectively.

TRA is one of the most prominent theories of social psychology and the most powerful to predict a specific behavior (Agarwal, 1999; Liska, 1984). Two meta-analyses of empirical studies testing the TRA conducted by Sheppard, Hartwick, and Warshaw (1988) show that the theory has a powerful predictive utility. Moreover, it is proved to be broadly applicable to a large array of behaviors (Agarwal, 1999; Sheppard et al., 1988). A study of a large variety of behaviors, such as voting for a presidential candidate, having an abortion, or smoking marijuana, confirms the causal relationships of the model, considering attitudes and subjective norms as independent variables and intention as the dependent variable (Ajzen, 1988).

From an empirical point of view, several researches applied TAM to different kinds of IT, such as text processing application (Davis et al., 1989), computers (Igbaria, 1995), e-mail (Karahanna & Straub, 1999), the World Wide Web (Moon & Kim, 2001), with different kinds of end user populations, such as students (Davis et al., 1989; Mathieson, 1991; Szajna, 1996), employees (Karahanna, Straub, & Chervany, 1999; Lucas & Spitler, 2000), or even physicians (Hu, Chau, Sheng, & Tam, 1999), and in different organizational contexts. Other studies test TAM and compare it to competing theories such as TRA (Davis et al., 1989) or Theory of Planned Behavior (Mathieson, 1991). All of these studies show that TAM has a strong predictive power. Moreover, TAM extensions (Agarwal & Karahanna, 2000; Agarwal & Prasad, 1999; Dishaw & Strong, 1999; Lucas & Spitler, 1999; Venkatesh & Davis, 2000; Venkatesh, Speir, & Morris, 2002), the main objective of which is to extend TAM by integrating other variables hypothesized to influence directly TAM constructs or indirectly by moderating the relationships among them, support its causal relationships and confirm its predictive utility. A nonexhaustive summary of TAM extensions is presented in Table 4.

From a practical point of view, TAM gives the possibility to practitioners, especially IS designers, to identify the areas in which we can act to improve IT use. Indeed, the two key concepts of TAM (i.e., perceived usefulness and perceived ease of use) could be controlled by IT designers. Consequently, developers could make it easy to use IT and to identify more concretely managerial benefits driven by IT that they are developing.

### THE INFLUENCE OF NATIONAL CULTURE ON IT ADOPTION: CONCEPTUAL MODEL AND RESEARCH PROPOSITIONS

We draw upon an extension of the TAM that integrates subjective norms (Venkatesh & Morris, 2000) as a determinant of the intention to use IT. Subjective norms reflect a person’s perception of the social pressure to perform the behavior in question. The importance attributed to social norms in determining and predicting behavior varies across cultures (Triandis, 1977); therefore, we expect that integrating subjective norms will strengthen our understanding of differences in behavioral intentions and will allow a better capturing of the cultural effect on IT use.

In order to explore the cultural factor, we rely upon Hofstede’s four cultural dimensions:
### Table 4. Summary of TAM extensions

<table>
<thead>
<tr>
<th>Authors</th>
<th>IT Tested</th>
<th>Variables Added to TAM</th>
<th>Research Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taylor and Todd (1995)</td>
<td>Hardware and software in a computing resource center</td>
<td>- Subjective norms&lt;br&gt;- Perceived behavioral control</td>
<td>Field study&lt;br&gt;Questionnaire addressed to 1,000 students from a business school</td>
</tr>
<tr>
<td>Igbaria (1995)</td>
<td>Computers</td>
<td>- Subjective Norms&lt;br&gt;- Normative beliefs and motivation to comply with the referent group&lt;br&gt;- Computer anxiety&lt;br&gt;- Computer knowledge&lt;br&gt;- Direction support&lt;br&gt;- Info center support&lt;br&gt;- Organizational politics&lt;br&gt;- Organizational use of the system (colleagues, subordinates, and CEOs)</td>
<td>Field study&lt;br&gt;Questionnaire addressed to 471 managers and professionals working in 54 North-American firms</td>
</tr>
<tr>
<td>Jackson, Chow, and Leitch (1997)</td>
<td>Different kinds of IS</td>
<td>- Situational involvement&lt;br&gt;- Intrinsic involvement&lt;br&gt;- Prior use</td>
<td>Filed study&lt;br&gt;Questionnaire addressed to 585 employees from organizations developing or revising their IS</td>
</tr>
<tr>
<td>Karahanna and Straub (1999)</td>
<td>Electronic Mail</td>
<td>- Social Influence&lt;br&gt;- Social Presence&lt;br&gt;- Perceived Accessibility&lt;br&gt;- Availability of User Training and Support</td>
<td>Field study&lt;br&gt;Questionnaire addressed to 100 users working in a worldwide American firm in the transportation sector</td>
</tr>
<tr>
<td>Venkatesh (1999)</td>
<td>Virtual Workplace System</td>
<td>- Game-based training</td>
<td>Experiment with business professionals: 69 in the first study and 146 in the second</td>
</tr>
<tr>
<td>Dishaw and Strong (1999)</td>
<td>Maintenance support software tools</td>
<td>- Technology-Task Fit (TTF)&lt;br&gt;- Task Characteristics&lt;br&gt;- Tool Experience&lt;br&gt;- Tool Functionality</td>
<td>Field study&lt;br&gt;Questionnaire addressed to programmer analysts working in three American firms, leaders in their respective fields (financial services, aerospace manufacturing, and insurance)</td>
</tr>
<tr>
<td>Lucas and Spitler (2000)</td>
<td>Broker workstations.</td>
<td>- Broker strategy (work approach)&lt;br&gt;- Perceived System Quality</td>
<td>Field study&lt;br&gt;Questionnaire addressed to 41 brokers</td>
</tr>
<tr>
<td>Venkatesh and Morris (2000)</td>
<td>System for data and information retrieval</td>
<td>- Gender&lt;br&gt;- Subjective Norms&lt;br&gt;- Experience</td>
<td>Longitudinal field study&lt;br&gt;Questionnaire addressed to 445 individuals from five organizations</td>
</tr>
<tr>
<td>Agarwal and Karahanna (2000)</td>
<td>World Wide Web</td>
<td>- Personal Innovativeness&lt;br&gt;- Playfulness&lt;br&gt;- Self-efficacy&lt;br&gt;- Cognitive Absorption</td>
<td>Field study&lt;br&gt;Questionnaire addressed to 250 students</td>
</tr>
<tr>
<td>Venkatesh and Davis (2000)</td>
<td>- Scheduling information system&lt;br&gt;- Windows-based environment&lt;br&gt;- Windows-based customer account management system</td>
<td>- Subjective norms&lt;br&gt;- Image&lt;br&gt;- Job relevance&lt;br&gt;- Output quality&lt;br&gt;- Result demonstrability&lt;br&gt;- Experience&lt;br&gt;- Voluntariness</td>
<td>Three longitudinal field studies&lt;br&gt;Questionnaires addressed to 48 floor supervisors in a medium-sized manufacturing firm in the first study, to 50 members of the financial department of a large financial firm in the second study, and to 51 employees from a small accounting services firm</td>
</tr>
<tr>
<td>Moon and Kim (2001)</td>
<td>World Wide Web</td>
<td>- Perceived Playfulness</td>
<td>Field study&lt;br&gt;Questionnaire addressed to 152 students</td>
</tr>
</tbody>
</table>
Information Technology Acceptance across Cultures

- **Individualism/collectivism:** Defined as the degree to which people in a society are integrated into strong cohesive ingroups.
- **Masculinity/femininity:** Defined as the extent to which a society attributes qualities such as assertiveness, material success to men, and modesty and concern about the quality of life to women.
- **Uncertainty avoidance:** Defined as the extent to which the members of a culture feel threatened by uncertain and unknown situations.
- **Power distance:** Defined as the extent to which the less powerful members of institutions and organization within a country expect and accept that power is distributed unequally.

Moreover, we rely upon Hall’s work on intercultural communication, which is based on two key cultural dimensions:

- **High context/low context:** Defined as the extent to which people are aware of and pay attention to all types of situational and contextual cues when interpreting messages.
- **Polychronism (vs. monochronism):** Defined as the preference to perform many activities at the same time.

These dimensions are hypothesized to have direct effects or moderator ones on TAM constructs and relationships. Integrating cultural dimensions to TAM is an attempt to better understand the genuine influence of national culture on IT adoption at the individual level. Indeed, cross-cultural studies on IS (see Table 1) focused on IS phenomena at the organizational level, such as IT transfer. Moreover, all the authors (except Srite & Karahanna, 2006) do not conceptualize national culture at the individual level. They have just studied IT-related behaviors in different countries, supposing that each country is characterized by a different set of cultural values. Their findings can be criticized in a sense that cultural values that characterize a country could not be espoused by all the members of this country. For example, Hofstede has found that Japan is a collectivist country, but this does not mean that each Japanese person will have the same level of collectivism. Actually, Hofstede has studied work-related values at the macro level by developing national scores for each cultural value. Therefore, it is worthwhile to analyze and measure culturally espoused values when dealing with national culture at the individual level. In this study, since we are attempting to deepen our understanding of IT adoption at the individual level, we have chosen to conceptualize cultural dimensions at the individual level so that they can be integrated in TAM. The conceptual model driven by this integration is presented in Figure 1.

In the next paragraphs, we present the rationale underlying model propositions.

**Proposition 1.** Individuals espousing high-context values will have a less strong perception of IT usefulness than those holding low-context values.

Since high-context oriented individuals prefer less coded messages, the electronic communication, implying very explicit content, will not be very useful as a means of communication. It cannot inform the receiver of the context of the message. The receiver probably will seek to know more about the message by other traditional means, implying more physical cues (face expression, intonation of the voice). Martinsons and Westwood (1997) have demonstrated that the use of management information systems (MIS) in Chinese culture depends on high-context communications. “Computers do not convey the necessary richness of meaning in a high context communication environment” (Martinsons & Westwood, 1997, p. 220) Indeed, since in China, a high-context country, coding is very restricted and so messages are terse in words, using computer-based MIS will not allow a full understanding of
messages. Actually, in order to better interpret the words, the Chinese rely upon other cues such as tone, dynamics, facial expression, and body language, which are not permitted when using computer-based communication.

**Proposition 2.** Individuals adopting monochronic time values will be using IT more than those adopting polychronic time values.

Individuals adopting the monochronic time are worried about the monitoring of time and prefer doing one thing at a time; therefore, they can view in IT a way to have control over time. Rowe and Struck (1999) argued that “people who are biased towards action see in asynchronous media the way to coordinate their tasks without interrupting them” (p. 166). In this case, we could expect that e-mail or voicemail would be appreciated highly by individuals espousing monochronic cultural values. On the contrary, individuals holding more fluid attitudes toward time are able to engage in many things at once, pay more attention to interpersonal relationships than to schedules, and accept disturbances. Therefore, they will be expected to view IT as less useful compared to individuals holding monochronic values.

**Proposition 3.** Individuals in high power distance cultures accept less IT than individuals in low power distance cultures.

In high power distance cultures, IT could threaten the hierarchy, which reflects “the existential inequality between higher-ups and lower-downs” (Hofstede, 1997, p. 37), because it suggests decentralization. Conversely, in low power distance cultures, individuals are more interdependent, whatever their ranks in the hierarchy. Therefore, they will be more favorable to IT, which does not contradict their perception of power distribution. For example, in an interpretive study conducted by Hasan and Ditsa (1999) comparing IT adoption in Australia, West Africa, and the Middle East, they have found that in the Middle East, high in power distance, modern IT, like the Internet, was seen as a threat to the authoritarian structure of their society. In the same vein, Martinsons and Westwood (1997) have demonstrated through an explanatory theory that paternalism in Chinese culture is an influential factor regarding the use of computer-based MIS, compared to American culture.

Straub, Keil, and Brenner (1997), testing TAM in three countries (the U.S., Switzerland, and Ja-
Information Technology Acceptance across Cultures

Proposition 4. Masculinity/femininity moderates the relationship between perceived usefulness and system usage.

Perceived usefulness will be more determinate for system usage for individuals espousing masculine values, since the perceived usefulness of IT is related intimately to achievement values that predominate in masculine cultures. In feminine cultures, work environment and the quality of life are much more important than attaining goals; this is why we expect much less impact of perceived usefulness of IT on usage for individuals adopting feminine values. Leidner and Carlsson (1998) have found that the Mexicans, belonging to a masculine culture, show more interest in executive information systems (EIS) than the Swedes. More precisely, the Mexicans perceive EIS as more useful than the Swedes, because such a system allows increasing the speed of decision making, and therefore, work goals could be reached more easily. In addition, in masculine cultures in which individuals are highly assertive, media lacking social presence, such as e-mail, would not be favored (Straub et al., 1997), because it does not allow end users to be highly aggressive and assertive.

Proposition 5. Individualism/collectivism moderates the relationship between subjective norms and system usage.

People in individualist cultures are more concerned with themselves than with the group. Therefore, for individuals holding individualistic cultural values, the opinions of the members of the group will not have weight in their decision to adopt IT. Conversely, people holding collectivist values will be more concerned with the maintenance of the group's cohesiveness. As a result, they will be expected to show more interest in others' opinions about IT because they want to conform to the group's behavior. In a comparative study on the applicability of the TRA across two different cultural settings, Korea and the U.S., Lee and Green (1991) have found that social norms are more important for collectivistic cultures (Korea) than for individualistic ones (U.S.) in influencing behavior.

Proposition 6. Masculinity/femininity moderates the relationship between subjective norms and system usage, such as subjective norms, will have much less influence on system usage for masculinity-oriented individuals.

In feminine cultures, individuals are expected to pay more attention to the opinions of others in behaving since they are more people-oriented than in masculine cultures in which the most important thing is goal achievement.

Proposition 7. Power distance moderates the relationship between subjective norms and system usage.

In high power distance cultures, since individuals are not supposed to disagree with their superiors, their reliance upon the opinions of
superiors will be more marked when assessing IT than for individuals from low power distance cultures.

Martinsons and Westwood (1997) have demonstrated through their explanatory theory that paternalism in Chinese culture is an influential factor regarding the use of management information systems compared to American culture. Paternalism is a concept very close to high power distance, since in paternalistic societies, key decisions are made by the patriarch, and management systems are very centralized.

According to Kowner and Wiseman (2003), people in tight (vs. loose) cultures in which deviation from behavior is not permitted “are more likely to obey the behavior ascribed by their social position” (p. 182). Consequently, social influence will play a greater role in shaping IT adoption for individuals belonging to tight culture than for those belonging to loose cultures.

**Proposition 8. Uncertainty avoidance acts also as a moderator of the relationship between subjective norms and system usage.**

The social influence exerted by important persons will be much more important in determining IT use in cultures seeking to avoid uncertainty than in cultures comfortable with uncertainty. Actually, in order to deal with uncertainty and ambiguity, individuals in strong uncertainty avoidance cultures are very concerned by the establishment and respect for rules, and therefore, subjective norms will be more important as guidance to behavior than for individuals in weak uncertainty avoidance cultures that rely more on their proper competence to evaluate a situation. The process of compliance (Kelman, 1958), one of the basic processes underlying the relationship between subjective norms and behavior, will be more salient when individuals are holding high uncertainty avoidance values. Indeed, compliance supposes that the individual will conform because a social actor that is important to him or her (and having the power of reward and punishment) wants him or her to behave in a specific manner. Therefore, conforming to referent others by adopting IT will weaken the uncertainty pertaining to the nonadoption of IT.

**PRACTICAL IMPLICATIONS**

The framework we provide is an attempt to highlight the importance of culture when dealing with IT adoption in different national cultural settings. It should improve IT designers’ awareness of the importance of cultural values for IT acceptance in foreign countries. Actually, the framework we offer does not suggest prescriptions to be adopted by IT designers. Rather, it encourages the latter to take into account cultural values when designing IS and/or preparing IT implementation procedures for companies around the world. Actually, IT designers should be aware that implementation failure could be caused by cultural incompatibility. As a result, they should adapt their implementation tactics so that they will be more congruent with end users’ cultural systems. For example, in high power distance cultures, end users may be reluctant to use computer-mediated communication tools, because this use may lead them to contradict their superiors. Therefore, IT designers should adjust IT features to the social needs of end users.

Furthermore, the conceptual model we have proposed suggests that the key belief in TAM, which is the perceived usefulness of IT, has cultural antecedents. Indeed, the perceived usefulness of IT depends on the type of the context of communication (high or low) and on the type of time adopted (monochronic or polychronic). In addition, the well-established positive relationship between perceived usefulness and system usage is shown to be moderated by two cultural values; namely, power distance and masculinity/femininity. Finally, the intensity of the influence of subjective norms on system usage has been
shown to be different, depending on the cultural values espoused by the IT users. Indeed, social influence would be most salient when end users accept and expect inequality in power, do not tolerate uncertainty, are more concerned by the group than by their selves, and are people-oriented.

CONCLUSION

In this chapter, we offer a theoretical perspective on the effect of national culture on IT adoption and use. Through the conceptual model we have presented, we wanted to move beyond models considering culture as a black box and offering an ad hoc cultural explanation of IT-related behaviors. First, we have identified pertinent cultural values that could offer a fruitful explanation of differences in IT-related behaviors and beliefs across countries. Then, we have provided propositions linking these cultural values to TAM constructs and relationships. We contend that high context/low context of communication and polychronism/monochronism have a direct effect on perceived usefulness of IT; power distance and masculinity/femininity have moderator effects on the relationship between perceived usefulness and system usage and individualism/collectivism; uncertainty avoidance, masculinity/femininity, and power distance have moderator effects on the relationship between subjective norms and system usage.

In the current context of internationalization, taking into account the societal environment in the understanding of IT acceptance behavior becomes more relevant in comparison with the past. Indeed, information technologies are so ubiquitous and pervasive that one should explore more their relationship with different societal contexts varying across countries. Actually, IT that is accepted in a specific cultural context may not be accepted in the same way in another culture.

The framework presented here is a first step in better understanding the relationship between national culture and IT adoption and use. It goes without saying that this conceptual model needs to be empirically tested. A field study research involving samples from different countries should be conducted.

REFERENCES


Information Technology Acceptance across Cultures

Israel and Taiwan. *Information and Management, 30*(1), 1-13.


---

*This work was previously published in Information Resources Management: Global Challenges, edited by W. Law, pp. 25-53, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 1.12
Is Information Ethics Culture–Relative?

Philip Brey
University of Twente, The Netherlands

ABSTRACT
In this article, I examine whether information ethics is culture relative. If it is, different approaches to information ethics are required in different cultures and societies. This would have major implications for the current, predominantly Western approach to information ethics. If it is not, there must be concepts and principles of information ethics that have universal validity. What would they be? The descriptive evidence is for the cultural relativity of information ethics will be studied by examining cultural differences between ethical attitudes towards privacy, freedom of information, and intellectual property rights in Western and non-Western cultures. I then analyze what the implications of these findings are for the metaethical question of whether moral claims must be justified differently in different cultures. Finally, I evaluate what the implications are for the practice of information ethics in a cross-cultural context.

INTRODUCTION
Information ethics has so far mainly been a topic of research and debate in Western countries, and has mainly been studied by Western scholars. There is, however, increasing interest in information ethics in non-Western countries like Japan, China, and India, and there have been recent attempts to raise cross-cultural issues in information ethics (e.g., Ess, 2002; Gorniak-Kocikowska, 1996; Mizutani, Dorsey & Moor, 2004). Interactions between scholars of Western and non-Western countries have brought significant differences to light between the way in which they approach issues in information ethics. This raises the question whether different cultures require a different information ethics and whether concepts and approaches in Western information ethics can be validly applied to the moral dilemmas of non-Western cultures. In other words, is information ethics culturally relative or are there concepts and principles of information ethics that have universal validity? The aim of this essay is to arrive at preliminary answers to this question.
MORAL RELATIVISM AND INFORMATION ETHICS

In discussions of moral relativism, a distinction is commonly made between descriptive and metaethical moral relativism. Descriptive moral relativism is the position that as a matter of empirical fact, there is extensive diversity between the values and moral principles of societies, groups, cultures, historical periods, or individuals. Existing differences in moral values, it is claimed, are not superficial but profound, and extend to core moral values and principles. Descriptive moral relativism is an empirical thesis that can in principle be supported or refuted through psychological, sociological, and anthropological investigations. The opposite of descriptive moral relativism is descriptive moral absolutism, the thesis that there are no profound moral disagreements exist between societies, groups, cultures, or individuals. At issue in this essay will be a specific version of descriptive moral relativism, descriptive cultural relativism, according to which there are major differences between the moral principles of different cultures.

Much more controversial than the thesis of descriptive moral relativism is the thesis of metaethical moral relativism, according to which the truth or justification of moral judgments is not absolute or objective, but relative to societies, groups, cultures, historical periods, or individuals. Whereas a descriptive relativist could make the empirical observation that one society, polygamy, is considered moral whereas in another it is considered immoral, a metaethical relativist could make the more far-reaching claim that the statement “polygamy is morally wrong” is true or justified in some societies while false or unjustified in others. Descriptive relativism therefore makes claims about the values that different people or societies actually have, whereas metaethical relativism makes claims about the values that they are justified in having. Metaethical moral relativism is antithetical to metaethical moral absolutism, the thesis that regardless of any existing differences between moral values in different cultures, societies, or individuals, there are moral principles that are absolute or objective, and that are universally true across cultures, societies, or individuals. Metaethical moral absolutism would therefore hold that the statement “polygamy is morally wrong” is either universally true or universally false; it cannot be true for some cultures or societies but false for others. If the statement is true, then societies that hold that polygamy is moral are in error, and if it is false, then the mistake lies with societies that condemn it.

The question being investigated in this essay is whether information ethics is culturally relative. In answering this question, it has to be kept in mind that the principal aims of information ethics are not descriptive, but normative and evaluative. That is, its principal aim is not to describe existing morality regarding information but rather to morally evaluate information practices and to prescribe and justify moral standards and principles for practices involving the production, consumption, or processing of information. A claim that information ethics is culturally relative therefore a claim that metaethical moral relativism is true for information ethics. It is to claim that the ethical values, principles, and judgments of information ethics are valid only relative to a particular culture, presumably the culture in which they have been developed. Since information ethics is largely a product of the West, an affirmation of the cultural relativity of information ethics means that its values and principles do not straightforwardly apply to non-Western cultures.

But if the cultural relativity of information ethics depends on the truth of metaethical relativism, does any consideration need to be given to descriptive relativism for information ethics? This question should be answered affirmatively. Defenses of metaethical relativism usually depend on previous observations that descriptive relativism is true. If descriptive relativism is false, it follows that people across the world share a moral
framework of basic values and principles. But if this is the case, then it seems pointless to argue for metaethical moral relativism: why claim that the truth of moral judgments is different for different groups if these groups already agree on basic moral values? On the other hand, if descriptive relativism is true, then attempts to declare particular moral principles of judgments to be universally valid come under scrutiny. Extensive justification would be required for any attempt to adopt a particular moral framework (say, Western information ethics) as one that is universally valid. In the next section, I will therefore focus on the question whether there are good reasons to believe that there are deep and widespread moral disagreements about central values and principles in information ethics across cultures, and whether therefore descriptive cultural relativism is true for information ethics.

THE DESCRIPTIVE CULTURAL RELATIVITY OF INFORMATION-RELATED VALUES

In this section, I will investigate the descriptive cultural relativity of three values that are the topic of many studies in information ethics: privacy, intellectual property, and freedom of information. Arguments have been made that these values are distinctly Western, and are not universally accepted across different cultures. In what follows I will investigate whether these claims seem warranted by empirical evidence. I will also relate the outcome of my investigations to discussions of more general differences between Western and non-Western systems of morality.

How can it be determined that cultures have fundamentally different value systems regarding notions like privacy and intellectual property? I propose that three kinds of evidence are relevant:

1. **Conceptual**: the extent to which there are moral concepts across cultures with similar meanings. For example, does Chinese culture have a concept of privacy that is similar to the American concept of privacy?

2. **Institutional**: the extent to which there is similarity between codified rules that express moral principles and codified statements that express moral judgments about particular (types of) situations. For example, are the moral principles exhibited in the laws and written rules employed in Latin cultures on the topic of privacy sufficiently similar to American laws and rules that it can be claimed that they embody similar moral principles?

3. **Behavioral**: the similarity between customs and behaviors that appear to be guided by moral principles. This would include tendencies to avoid behaviors that are immoral regarding a moral principle, tendencies to show disapproval to those who engage in such behaviors, and to show disapproval to those who do not, and tendencies to show remorse or guilt when engaging in such behaviors. For instance, if a culture has a shared privacy principle that states that peeking inside someone’s purse is wrong, then it can be expected that most people try not to do this, disapprove of those who do, and feel ashamed or remorseful when they are caught doing it.

It is conceivable that in a particular culture a value or moral principle is widely upheld at the behavioral level, but has not (yet) been codified at the institutional and conceptual level. But this is perhaps unlikely in cultures with institutions that include extensive systems of codified rules, which would include any culture with a modern legal system. It is also conceivable that a moral value or principle is embodied in both behavioral customs and codified rules, but no good match can be found at the conceptual level. In that case,
it seems reasonable to assume that the value or principle at issue is embodied in the culture, but different concepts are used to express it, making it difficult to find direct translations.

A full consideration of the evidence for descriptive moral relativism along these three lines is beyond the scope of this article. I only intend to consider enough evidence to arrive at a preliminary assessment of the cultural relativity of values in contemporary information ethics.

Privacy

It has been claimed that in Asian cultures like China and Japan, no genuine concept or value of privacy exists. These cultures have been held to value the collective over the individual. Privacy is an individual right, and such a right may not be recognized in a culture where collective interest tend to take priority over individual interests. Using the three criteria outline above, and drawing from studies of privacy in Japan, China and Thailand, I will now consider whether this conclusion is warranted.

At the conceptual level, there are words in Japanese, Chinese, and Thai that refer to a private sphere, but these words seem to have substantially different meanings than the English word for privacy. Mizutani et al. (2004) have argued that there is no word for “privacy” in traditional Japanese. Modern Japanese, they claim, sometimes adopt a Japanese translation for the Western word for privacy, which sounds like “puraibashii”, and written in katakana. Katakana is the Japanese phonetic syllabary that is mostly used for words of foreign origin. According to Nakada and Tamura (2005), Japanese does include a word for “private”, “Watakusi”, which means “partial, secret and selfish”. It is opposed to “Ohyake”, which means “public”. Things that are Watakusi are considered less worthy than things that are Ohyake. Mizutani et al. (2004) point out, in addition, that there are certainly behavioral customs in Japan that amount to a respect for privacy. There are conventions that restrict access to information, places, or objects. For example, one is not supposed to look under clothes on public streets.

In China, the word closest to the English “privacy” is “Yinsi”, which means “shameful secret” and is usually associated with negative, shameful things. Lü (2005) claims that only recently that “Yinsi” has also come to take broader meanings to include personal information, shameful or not, that people do not want others to know (see also Jingchun, 2005; McDougall & Hansson, 2002). This shift in meaning has occurred under Western influences. As for institutional encoding of privacy principles, Lü maintains that there currently are no laws in China that protect an individual right to privacy, and the legal protection of privacy has been weak and is still limited, though there have been improvements in privacy protection since the 1980s.

Kitiyadisai (2005), finally, holds that the concept of privacy does not exist in Thailand. She claims that the Western word privacy was adopted in the late nineteenth or early twentieth century in Thailand, being transliterated as “privade,” but this word gained a distinctly Thai meaning, being understood as a collectivist rather than an individual notion. It referred to a private sphere in which casual dress could be worn, as opposed to a public sphere in which respectable dress had to be worn. In the Thai legal system, Kitiyadisai claims there has not been any right to privacy since the introduction of privacy legislation in 1997 and a Thai constitution, also in 1997, that for the first time guarantees basic human rights. Kitiyadisai argues, however, that Thai privacy laws are hardly enacted in practice, and many Thais remain unaware of the notion of privacy.

It can be tentatively concluded that the introduction of a concept of privacy similar to the Western notion has only taken place recently in Japan, China, and Thailand, and that privacy legislation has only taken place recently. In traditional
Japanese, Chinese, and Thai culture, which still has a strong presence today, distinctions are made that resemble the Western distinction between public and private, and customs exist that may be interpreted as respective of privacy, but there is no recognized individual right to privacy.

**Intellectual Property Rights**

In discussing the cultural relativity of intellectual property rights (IPR), I will limit myself to one example: China. China is known for not having a developed notion of private or individual property. Under communist rule, the dominant notion of property was collective. All means of production, such as farms and factories, were to be collectively owned and operated. Moreover, the state exercised strict control over the means of production and over both the public and private sphere. A modern notion of private property was only introduced since the late 1980s. Milestones were a 1988 constitutional revision that allowed for private ownership of means of production and a 2004 constitutional amendment that protects citizens from encroachment of private property.

The notion of intellectual property has only recently been introduced in China, in the wake of China’s recent economic reforms and increased economic interaction with the West. China is currently passing IPR laws and cracking down on violations of IPR in order to harmonize the Chinese economic system with the rest of the world. But as journalist Ben Worthen observes, “the average citizen in China has no need and little regard for intellectual property. IPR is not something that people grew up with … and the percent of citizens who learn about it by engaging in international commerce is tiny.” Worthen also points out that Chinese companies “have no incentive to respect IPR unless they are doing work for Western companies that demand it” and that “since most of the intellectual property royalties are headed out of China there isn’t a lot of incentive for the government to crack down on companies that choose to ignore IPR.”

All in all, it can be concluded that China’s value system traditionally has not included a recognition of intellectual property rights, and it is currently struggling with this concept.

**Freedom of Information**

Freedom of information is often held to comprise two principles: freedom of speech (the freedom to express one’s opinions or ideas, in speech or in writing) and freedom of access to information. Sometimes, freedom of the press (the freedom to express oneself through publication and dissemination) is distinguished as a third principle. In Western countries, freedom of information is often defined as a constitutional and inalienable right. Laws protective of freedom of information are often especially designed to ensure that individuals can exercise this freedom without governmental interference or constraint. Government censorship or interference is only permitted in extreme situations, pertaining to such things as hate speech, libel, copyright violations, and information that could undermine national security.

In many non-Western countries, freedom of information is not a guiding principle. There are few institutionalized protections of freedom of information; there are many practices that interfere with freedom of information, and a concept of freedom of information is not part of the established discourse in society. In such societies, the national interest takes precedence, and an independent right to freedom information either is not recognized or is made so subordinate to national interests that it hardly resembles the Western right to freedom of information. These are countries in which practices of state censorship are widespread; mass media are largely or wholly government-controlled, the Internet, databases, and libraries are censored, and messages that do not conform to the party line are cracked down upon.

Let us, as an example, consider the extent to which freedom of information can be said to be a value in Chinese society. Until the 1980s, the
idea of individual rights or civil rights was not a
well-known concept in China. Government was
thought to exist to ensure a stable society and a
prosperous economy. It was not believed to have
a function to protect individual rights against col-
clective and state interests. As a consequence of
this general orientation, the idea of an individual
right to freedom of information was virtually
unknown. Only recently has China introduced
comprehensive civil rights legislation. In its 1982
constitution, China introduced constitutional
principles of freedom of speech and of the press.
And in 1997, it signed the International Conven-
tion on Economic, Social, and Cultural Rights,
and in 1998 the International Convention on Civil
and Political Rights (the latter of which it has not
yet ratified).

Even though the Chinese government has
recently come to recognize a right to freedom of
information, as well as individual human rights
in general, and has introduced legislation to this
effect, state censorship is still rampant, and the
principle of upholding state interest still tends to
dominate the principle of protecting individual
human rights. Internet censorship presents a good
example of this. Internet traffic in China is con-
trolled through what the Chinese call the Golden
Shield, and what is known outside mainland China
as the Great Firewall of China. This is a system
of control in which Internet content is blocked by
routers, as well as at the backbone and ISP level,
through the “filtering” of undesirable URLs and
keywords. A long list of such “forbidden” URLs
and keywords has been composed by the Chinese
State Council Information Office, in collaboration
with the Communist Party’s Propaganda Depart-
ment. This system is especially geared towards
censorship of content coming from outside main-
land China (Human Rights Watch, 2006).

Rights-Centered and
Virtue-Centered Morality

A recurring theme in the above three discussions
has been the absence of a strong tradition of individual
rights in the cultures that were discussed—those of China, Japan, and Thailand—and the
priority that is given to collective and state interests. Only very recently have China, Japan, and
Thailand introduced comprehensive human rights
legislation, which has occurred mainly through
Western influence, and there is still considerable
tension in these societies, especially in China
and Thailand, between values that prioritize the
collective and the state and values that prioritize
the individual.

Various authors have attempted to explain the
worldview that underlies the value system of these
countries. In Japan and Thailand, and to a lesser
extent China, Buddhism is key to an understanding
of attitudes towards individual rights. Buddhism
holds a conception of the self that is antithetical
to the Western conception of an autonomous
self which aspires to self-realization. Buddhism
holds that the self does not exist and that human
desires are delusional. The highest state that hu-
mans can reach is Nirvana, a state of peace and
contentment in which all suffering has ended. To
reach Nirvana, humans have to become detached
from their desires, and realize that the notion of
an integrated and permanent self is an illusion. In
Buddhism, the self is defined as fluid, situation-
dependent, and ever-changing. As Mizutani et al.
and Kitiyadisai have noted, such a notion of the
self is at odds with a Western notion of privacy
and of human rights in general, notions which
presuppose a situation-independent, autonomous
self which pursues its own self-interests and which
has inalienable rights that have to be defended
against external threats.

In part through Buddhism, but also through
the influence of other systems of belief such as
Confucianism, Taoism, and Maoism, societies
Is Information Ethics Culture-Relative?

like those of China and Thailand have developed a value system in which the rights or interests of the individual are subordinate to those of the collective and the state. To do good is to further the interests of the collective. Such furtherances of collective interests will generally also benefit the individual. The task of government, then, is to ensure that society as a whole functions well, in a harmonious and orderly way, and that social ills are cured, rather than the ills of single individuals. In other words, government works for the common good, and not for the individual good.

Only recently have countries like China and Thailand come to recognize individual human rights and individual interests next to collective interests. But according to Lü (2005), the collectivist ethic still prevails:

Adapting to the demands of social diversity, the predominant ethics now express a new viewpoint that argues against the simple denial of individual interests and emphasizes instead the dialectical unification of collective interests and individual interests: in doing so, however, this ethics points out that this kind of unification must take collective interests as the foundation. That is to say, in the light of the collectivism principle of the prevailing ethics, collective interests and individual interests are both important, but comparatively speaking, the collective interests are more important than individual interests. (Lü, 2005, p. 12)

If this observation is correct, then the introduction of human rights legislation and property rights in countries like China is perhaps not motivated by a genuine recognition of inalienable individual human rights, but rather a recognition that in the current international climate, it is better to introduce human rights and property rights, because such principles will lead to greater economic prosperity, which is ultimately to the benefit of the collective.

The dominant value systems prevalent in China, Thailand, and Japan are examples of what philosopher David Wong (1984) has called virtue-centered moralities. According to Wong, at least two different approaches to morality can be found in the world: a virtue-centered morality that emphasizes the good of the community, and a rights-centered morality that stresses the value of individual freedom. Rights-centered morality is the province of the modern West, although it is also establishing footholds in other parts of the world. Virtue-centered morality can be found in traditional cultures such as can be found in southern and eastern Asia and in Africa. Wong’s distinction corresponds with the frequently made distinction between individualist and collectivist culture, that is found, amongst others, in Geert Hofstede’s (1991) well-known five-dimensional model of cultural difference. However, this latter distinction focuses on social systems and cultural practices, whereas Wong makes a distinction based in differences in moral systems.

In Wong’s conception of virtue-centered moralities, individuals have duties and responsibilities that stem from the central value of a common good. The common good is conceived of in terms of an ideal conception of community life, which is based on a well-balanced social order in which every member of the community has different duties and different virtues to promote the common good. Some duties and virtues may be shared by all members. The idea that human beings have individual rights is difficult to maintain in this kind of value system, because recognition of such rights would have to find its basis in the higher ideal of the common good. But it seems clear that attributing rights to individuals is not always to the benefit of the common good. The recognition of individual property rights, for example, could result in individual property owners not sharing valuable resources that would benefit the whole community. In virtue-centered moralities, the ideal is for individuals to be virtuous, and virtuous individuals are those individuals whose individual good coincides with their contribution to the common good. Individual goods may be recog-
Is Information Ethics Culture-Relative?

Individuals deserve respect only because of their perceived contribution to the common good, not because they possess inalienable individual rights.

Conclusion

The discussion of privacy, intellectual property rights, and freedom of information has shown that a good case can be made for the descriptive cultural relativity of these values. These values are central in information ethics, as it has been developed in the West. Moreover, it was argued that the uncovered cultural differences in the appraisal of these values can be placed in the context of a dichotomy between two fundamentally different kinds of value systems that exist in different societies: rights-centered and virtue-centered systems of value. Information ethics, as it has developed in the West, has a strong emphasis on rights, and little attention is paid to the kinds of moral concerns that may exist in virtue-centered systems of morality. In sum, it seems that the values that are of central concern in Western information ethics are not the values that are central in many non-Western systems of morality. The conclusion therefore seems warranted that descriptive moral relativism is true for information ethics.

Two Standard Arguments for Metaethical Relativism

There are two traditional arguments for metaethical moral relativism that rely on the truth of descriptive moral relativism (Wong, 1993). The one most frequently alluded to is the argument from diversity. This argument starts with the observation that different cultures employ widely different moral standards. Without introducing additional premises, the argument goes on to conclude that therefore, there are no universal moral standards. This argument rests on what is known in philosophy as a naturalistic fallacy, an attempt to derive a norm from a fact, or an “ought” from an “is”. The premise of the argument is descriptive: there are different moral standards. The conclusion is normative: no moral standard has universal validity. No evidence has been presented that the truth of the premise has any bearing on the truth of the conclusion.

A second, stronger argument for moral relativism is the argument from functional necessity, according to which certain ethical beliefs in a society may be so central to its functioning that they cannot be given up without destroying the society. Consequently, the argument runs, these ethical beliefs are true for that society, but not necessarily in another. However, this argument is also problematic because it grounds the truth of ethical statements in their practical value for maintaining social order in a particular society. Such a standard of justification for ethical statements is clearly too narrow, as it could be used to justify the moral beliefs of societies whose beliefs and practices are clearly unethical, for instance, fascist societies. If a society operates in a fundamentally unethical way, then the transformation of some of its social structures and cultural forms would seem acceptable if more ethical practices are the result.
Wong’s and Harman’s Argument for
Metaethical Relativism

More convincing arguments for moral relativism have been presented by David Wong (1984, 2006) and Gilbert Harman (1996, 2000). Their argument runs, in broad outline, as follows. There are deep-seated differences in moral belief between different cultures. Careful consideration of the reasons for these moral beliefs they have shown that they are elements of different strategies to realize related but different conceptions of the Good. No good arguments can be given why one of these conceptions of the Good is significantly better than all the others. Therefore, these moral beliefs are best explained as different but (roughly) equally valid strategies for attaining the Good.

This is a much better argument than the previous two, since it puts the ball in the metaethical absolutist’s court: he will have to come up with proof that it is possible to provide good arguments for the superiority of one particular conception of the Good over all other conceptions. Metaethical absolutists can respond to this challenge in two ways. First, they may choose to bite the bullet and claim that a rational comparison of different conceptions of the Good is indeed possible. Different conceptions of the Good, they may argue, rely on factual or logical presuppositions that may be shown to be false. Alternatively, they may argue that there are universally shared moral intuitions about what is good, and these intuitions can be appealed to in defending or discrediting particular conceptions of the Good. For instance an individual who believes that physical pleasure is the highest good could conceivably be persuaded to abandon this belief through exposure to arguments that purport to demonstrate that there are other goods overlooked by this individual that are at least as valuable. Such an argument could conceivably rely on someone’s moral intuitions about the Good that could be shown to deviate from someone’s explicit concept of the Good.

Second, a mixed position could be proposed, according to which it is conceded that individuals or cultures may hold different conceptions of the Good that cannot be rationally criticized (pace metaethical relativism) but that rational criticism of individual moral beliefs is nevertheless possible (pace metaethical absolutism) because these beliefs can be evaluated for their effectiveness in realizing the Good in which service they stand. After all, if moral beliefs are strategies to realize a particular conception of the Good, as Wong and Harman have argued, then they can be suboptimal in doing so. A belief that Internet censorship is justified because it contributes to a more stable and orderly society can be wrong because it may not in fact contribute to a more stable and orderly society. Empirical arguments may be made that Internet censorship is not necessary for the maintenance of social order, or even that Internet censorship may ultimately work to undermine social order, for example, because it creates discontentment and resistance.

In the existing dialogue between proponents of rights-centered and virtue-centered systems of morality, it appears that both these approaches are already being taken. Western scholars have criticized the organicist conception of society that underlies conceptions of the Good in many Asian cultures, while Western definitions of the Good in terms of individual well-being have been criticized for their atomistic conception of individuals. Rights-based systems of morality have been criticized for undervaluing the common good, whereas virtue-based systems have been criticized for overlooking the importance of the individual good. In addition, both rights-centered and virtue-centered systems of morality have been criticized for not being successful by their own standards. Western individualism has been claimed to promote selfishness and strife, which results in many unhappy individuals plagued by avarice, poverty, depression, and loneliness. Western societies have therefore been claimed to be unsuccessful in attaining their own notion of the Good, defined in terms of individual well-be-
Is Information Ethics Culture-Relative?

Virtue-centered cultures have been claimed to have difficulty in developing strong economies that serve the common good, because good economies have been argued to require private enterprise and a more individualist culture. In addition, strong state control, which is a feature of many virtue-centered cultures, has been argued to lead to corruption and totalitarianism, which also do not serve the common good.

In light of the preceding observations, it seems warranted to conclude, pace metaethical absolutism, that rational criticism between different moral systems is possible. It does not follow, however, that conclusive arguments for universal moral truths or the superiority of one particular moral system over others are going to be possible. Critics of a particular moral system may succeed in convincing its adherents that the system has its flaws and needs to be modified, but it could well be that no amount of criticism ever succeeds in convincing its adherents to abandon core moral beliefs within that system, however rational and open-minded these adherents are in listening to such criticism.

Conclusion

I have argued, pace metaethical relativism, that it is difficult if not impossible to provide compelling arguments for the superiority of different notions of the Good that are central in different moral systems, and by implication, that it is difficult to present conclusive arguments for the universal truth of particular moral principles and beliefs. I have also argued, pace metaethical absolutism, that is nevertheless possible to develop rational arguments for and against particular moral values and overarching conceptions of the Good across moral systems, even if such arguments do not result in proofs of the superiority of one particular moral system or moral principle over another.

From these two metaethical claims, a normative position can be derived concerning the way in which cross-cultural ethics ought to take place. It follows, first of all, that it is only justified for proponents of a particular moral value or principle to claim that it ought to be accepted in another culture if they make this claim on the basis of a thorough understanding of the moral system operative in this other culture. The proponent would have to understand how this moral system functions and what notion of the Good it services, and would have to have strong arguments that either the exogenous value would be a good addition to the moral system in helping to bring about the Good serviced in that moral system, or that the notion of the Good serviced in that culture is flawed and requires revisions. In the next section, I will consider implications of this position for the practice of information ethics in cross-cultural settings.

INFORMATION ETHICS IN A CROSS-CULTURAL CONTEXT

It is an outcome of the preceding sections that significant differences exist between moral systems of different cultures, that these differences have important implications for moral attitudes towards uses of information and information technology, and that there are good reasons to take such differences seriously in normative studies in information ethics. In this section, I will argue, following Rafael Capurro, that we need an intercultural information ethics that studies and evaluates cultural differences in moral attitudes towards information and information technology. I will also critically evaluate the claim that the Internet will enable a new global ethic that provides a unified moral framework for all cultures.

Intercultural Information Ethics

The notion of an intercultural information ethics (IIE) was first introduced by Rafael Capurro (2005, in press), who defined it as a field of research in which moral questions regarding information
technology and the use of information are reflected on in a comparative manner on the basis of different cultural traditions. I will adopt Capurro’s definition, but differ with him on what the central tasks of an IIE should be. Capurro defines the tasks of IIE very broadly. For him, they not only the comparative study of value systems in different cultures in relation to their use of information and information technology, but also studies of the effect of information technology on customs, languages, and everyday problems, the changes produced by the Internet on traditional media, and the economic impact of the Internet to the extent that it can become an instrument of cultural oppression and colonialism.

I hold, in contrast, that studies of the effects of information technology in non-Western cultures are more appropriately delegated to the social sciences (including communication studies, cultural studies, anthropology and science, and technology studies). An intercultural information ethics should primarily focus on the comparative study of moral systems. Its overall aim would be to interpret, compare, and critically evaluate moral systems in different cultures in relation to their moral attitudes towards and behavior towards information and information technology.

This task for IIE can be broken down into four subtasks, the first two of which are exercises in descriptive ethics and the latter two of which belong to normative ethics. First, IIE should engage in interpretive studies of moral systems in particular cultures, including the systems of value contained in the religious and political ideologies that are dominant in these cultures. The primary focus in such interpretive studies within the context of IIE should be on resulting moral attitudes towards the use and implications of information technology and on the moral problems generated by uses of information technology within the context of the prevailing moral system. Second, IIE should engage in comparative studies of moral systems from different cultures, and arrive at analyses of both similarities and differences in the way that these moral systems are organized and operate, with a specific focus on the way in which they have different moral attitudes towards implications of information technology and on differences in moral problems generated by the use of information technology.

Third, IIE should engage in critical studies in which the moral systems of particular cultures are criticized based on the insights gained through the interpretive and comparative studies alluded to above, particularly in their dealings with information technology. Critical studies may be directed towards criticizing moral values and beliefs in cultures other than one’s own, and proposing modifications in the culture’s moral system and ways in which it should solve moral problems, but may also involve self-criticism, in which one’s own moral values and the moral system of one’s own culture is criticized based on insights gained from the study of alternative moral systems. Fourth, IIE should engage in interrelational studies that focus on the construction of normative models for interaction between cultures in their dealings with information and information technology that respect their different moral systems. Interrelational studies hence investigate what moral compromises cultures can make and ought to make in their interactions and what shared moral principles can be constructed to govern their interactions.

Global Ethics and the Information Revolution

Some authors have argued that globalization and the emergence of the Internet have created a global community, and that this community requires its own moral system that transcends and unifies the moral systems of all cultures and nations that participate in this global community. The ethics needed for the construction of such a moral system has been called global ethics. The idea of a global ethics or ethic was first introduced by German theologian Hans Küng in 1990 and later
elaborated by him in a book (Küng, 2001). His aim was to work towards a shared moral framework for humanity that would contain a minimal consensus concerning binding values and moral principles that could be invoked by members of a global community in order to overcome differences and avoid conflict.

Krystyna Górniak-Kocikowska (1996) has argued that the computer revolution that has taken place has made it clear that a future global ethic will have to be a computer ethic or information ethic. As she explains, actions in cyberspace are not local, and therefore the ethical rules governing such actions cannot be rooted in a particular local culture. Therefore, unifying ethical rules have to be constructed in cyberspace that can serve as a new global ethic. Similar arguments have been presented by Bao and Xiang (2006) and De George (2006).

No one would deny that a global ethic, as proposed by Küng, would be desirable. The construction of an explicit, shared moral framework that would bind all nations and cultures would evidently be immensely valuable. It should be obvious, however, that such a framework could only develop as an addition to existing local moral systems, not as a replacement of them. It would be a framework designed to help solve global problems, and would exist next to the local moral systems that people use to solve their local problems. In addition, it remains to be seen if cross-cultural interactions over the Internet yield more than a mere set of rules for conduct online, a global netiquette, and will result in a global ethic that can serve as a common moral framework for intercultural dialogue and joint action. Hongladarom (2001) has concluded, based on empirical studies, that the Internet does not create a worldwide monolithic culture but rather reduplicates existing cultural boundaries. It does create an umbrella cosmopolitan culture to some extent, but only for those Internet users who engage in cross-cultural dialogue, which is a minority, and this umbrella culture is rather superficial. Claims that the Internet will enable a new global ethic may therefore be somewhat premature. In any case, such intercultural dialogue online will have to be supplemented with serious academic work in intercultural information ethics, as well as intercultural ethics at large.

**CONCLUSION**

It was found in this essay that very different moral attitudes exist in Western and non-Western countries regarding three key issues in information ethics: privacy, intellectual property, and freedom of information. In non-Western countries like China, Japan, and Thailand, there is no strong recognition of individual rights in relation to these three issues. These differences were analyzed in the context of a difference, proposed by philosopher David Wong, between rights-centered moralities that dominate in the West and virtue-centered moralities that prevail in traditional cultures, including those in South and East Asia. It was then argued that cross-cultural normative ethics cannot be practiced without a thorough understanding of the prevailing moral system in the culture that is being addressed. When such an understanding has been attained, scholars can proceed to engage in moral criticism of practices in the culture and propose standards and solutions to moral problems. It was argued, following Rafael Capurro, that we need an intercultural information ethics that engages in interpretive, comparative, and normative studies of moral problems and issues in information ethics in different cultures. It is to be hoped that researchers in both Western and non-Western countries will take up this challenge and engage in collaborative studies and dialogue on an issue that may be of key importance to future international relations.
REFERENCES


Is Information Ethics Culture-Relative?


ENDNOTES

1 By information ethics I mean the study of ethical issues in the use of information and information technology. Contemporary information ethics is a result of the digital revolution (or information revolution) and focuses mainly on ethical issues in the production, use, and dissemination of digital information and information technologies. It encloses the field of computer ethics (Johnson, 2000) as well as concerns that belong to classical information ethics (which was a branch of library and information science), media ethics, and journalism ethics.

2 This doctrine is called metaethical rather than normative because it does not make any normative claims, but rather makes claims about the nature of moral judgments. *Normative moral relativism* would be the thesis that it is morally wrong to judge or interfere with the moral practices of societies, groups, cultures, or individuals who have moral values different from one’s own. This is a normative thesis because it makes prescriptions for behavior.


Chapter 1.13
Personalization Techniques and Their Application

Juergen Anke
Dresden University of Technology, Germany

David Sundaram
University of Auckland, New Zealand

INTRODUCTION

Personalization is an approach to increase the usability of complex information systems and present the user with a comprehensible interface that is tailored to his or her needs and interests. In this article, we examine general techniques that are employed to achieve the personalization of Web sites. This is followed by a presentation of real-world examples. It will be shown how different levels of personalization can be achieved by employing the discussed techniques. This leads finally to a summary of the current state in personalization technologies and the issues connected with them. The article closes with some ideas on further research and development, and a conclusion.

In general, the concept of personalization refers to the ability of tailoring standardized items to the needs of individual people. It is originally derived from the ideas of Pine (1993) who proposed that companies should move from the paradigms of standardized products and homogeneous markets to customizable products that meet the requirements of many different customers. The principle of mass customization applies to a certain degree to most car manufacturers and some computer manufacturers, for example, Dell.

In the digital world of the World Wide Web, the degree of customization can be much higher than in the physical world. Currently, a number of online portals and e-commerce shops make use of personalization to provide a better user experience. Although Web sites may be the most popular examples of personalization, the concept is not limited to the Web. Every information system that deals with large amounts of data and/or has a heterogeneous group of users can benefit from it. Examples include e-learning environments, electronic books, computer-operated voice and telephony services, and tourist guides.

Personalization is also very useful for mobile devices like personal digital assistants (PDAs) or mobile phones (cf, Mulvenna, Anand, & Buchner,
2000). Technologies like mobile Internet access, WAP (Wireless Application Protocol), and future multimedia applications based on high-capacity wireless technologies require the designers of services for these devices to deal with limited input capabilities and small display sizes. For that reason, every method that assists the user in navigating and finding information easily adds real value to applications for such devices.

PERSONALIZATION TECHNIQUES

The core idea of personalization is to customize the presentation of information specifically to the user to make user interfaces more intuitive and easier to understand, and to reduce information overload.

The main areas of tailoring presentation to individual users are content and navigation. Content refers to the information being displayed, and navigation refers to the structure of the links that allow the user to move from one page to another. Personalized navigation can help the user to easily find what he or she is looking for or to discover new information. For example, a system discussed by Belkin (2000) assists users in refining search queries by giving recommendations on related or similar terms.

Adaptable vs. Adaptive

There are two approaches to achieve personalization: adaptable and adaptive methods. The former is a term for systems that can be customized by the user in an explicit manner; that is, the user can change the content, layout, appearance, and so forth to his or her needs. This data is called a user profile, and all personalized presentation is based on data the user provided for configuration purposes. It is important to note that the customized appearance does not change over time until the user decides to change his or her preferences.

In contrast, adaptive methods change the presentation implicitly by using secondary data. This data can be obtained from a variety of sources, for example, from the user’s actions, from the behaviour of other users on that site, or based on the currently displayed content. Methods that use this data as input are discussed in detail below. The most distinctive characteristic of adaptive methods is that they are constantly monitoring the user’s activities to adjust the arrangement and selection of relevant information.

Adaptive methods or machine-learning algorithms are huge steps toward automated customization. Current static interfaces suffer from the fact that the designer has to anticipate the needs, interests, and previous knowledge of the users in advance. As these preferences change over time, customization that requires human interaction for collecting and identifying preferences leads quickly to outdated user profiles.

Table 1 shows how adaptive and adaptable methods can be applied to customize content and navigation. The examples given are intended to be generic; more concrete examples are examined in the case studies below.

Degree of Personalization

Another important criterion for classification is the degree of personalization. Systems can have transient or persistent personalization, or be non-persnalized. With transient personalization, the customization remains temporary and is largely based on a combination of the user’s navigation and an item-to-item correlation. For example, if an item is selected, the system attaches similar items as recommendations to it whereby the content of the shopping cart is taken into consideration.

Persistent personalization systems maintain a permanent user account for every user to preserve his or her settings and preferences across separate sessions. Although this raises privacy issues and is the most difficult to implement, it offers the greatest benefit. These systems can make use
of user-to-user correlation algorithms and thus provide higher accuracy.

Another technology that belongs to the broad area of personalization is recommender systems (Mulvenna et al., 2000). Whereas straight personalization tailors just the presentation of information, recommender systems support the user in discovering new information. As recommendation relies on user preferences and interests, it is often part of personalized systems. From another perspective, one can say that recommender systems provide a selection of the most suitable content for the user. The application of recommender systems to e-commerce is discussed by Schafer, Konstan, and Riedl (2001).

In the following sections, we look at two cases that highlight the adaptive and adaptable approaches to personalization.

<table>
<thead>
<tr>
<th>Adaptable</th>
<th>Content</th>
<th>Navigation</th>
</tr>
</thead>
<tbody>
<tr>
<td>- explicit selection and ordering of content items by the user</td>
<td>- building link lists (favourites, bookmarks)</td>
<td></td>
</tr>
<tr>
<td>- providing personal information to be listed in directories</td>
<td>- setting default links for generic navigational structures/menus to omit intermediate step(s)</td>
<td></td>
</tr>
<tr>
<td>- setting up stock portfolios</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adaptive</th>
<th>Content</th>
<th>Navigation</th>
</tr>
</thead>
<tbody>
<tr>
<td>- present the user new items which are related to the current items (recommendations)</td>
<td>- hiding unsuitable links based on the context</td>
<td></td>
</tr>
<tr>
<td>- filter content based on current actions (remove items which are dissimilar)</td>
<td>- annotate links to give meta-information about value of the linked content relating to the user’s navigation history (e.g. “no new information”, “insufficient previous knowledge” etc.)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Application of adaptable and adaptive methods to content and navigation

APPLICATION AND IMPACT OF PERSONALIZATION: AMAZON.COM

Amazon.com is one of the pioneers of e-commerce. Originally set up as a bookstore, it has grown to a general retailer for a wide range of products. It also provides an auction platform and a marketplace where customers can sell used goods. The marketplace is seamlessly integrated into the main product catalogue, therefore customers can decide whether they want to buy a particular product as a new or a used one.

Goal

As the Amazon.com product catalogue contains more than 2 million products, users can easily get frustrated if they do not find what they are looking for. Thus, one of the main goals is to tailor the product catalogue as much as possible to the needs and interests of the user.

Aside from easy navigation, the site offers a seamlessly integrated recommendation system. It is intended to offer customers products that are either related to their interests or to the product that is currently displayed to exploit cross-selling potentials.

Personalization Techniques

Amazon.com is a highly developed online shopping site and incorporates a combination of
numerous adaptive and adaptable methods. The generated user profile is stored in a database on the server; that is, the degree of personalization is persistent.

The prevalent recommendation method is based on the purchases of other customers. It appears as a list beginning with the phrase “Customers who bought this book also bought…” on each product detail page (Figure 1a). A second list contains up to five authors whose books were bought by customers who also bought the currently selected book. Both of these features are collaborative-filtering methods of implicit collected data. Another feature called purchase circles is also a collaborative-filtering mechanism that displays the top-10 products for a particular region, institution, or company.

Other recommendation methods are based on the activities of the customer. For example, there is always one product recommendation in a sidebar that is apparently directly derived from the list of recently viewed items (Figure 1c). Moreover, there is also a recommendation page for each of the main categories (books, software, etc.) that is completely based on the customer’s purchase history.

Amazon.com also encourages customers to rate their purchased products (Figure 1e). Ratings belong to the collaborative-filtering methods. However, in contrast to the list of products other customers bought, they require explicit interaction for the sake of personalization. The ratings can also be part of customer reviews and comments. Interestingly, it is even possible to rate the ratings; that is, customers can indicate whether they found the rating helpful or not (e.g., “3 of 5 people found the following review helpful”). This mechanism is used to display the most helpful comments first and let the least helpful comments move to secondary pages.

As mentioned before, one of the goals is to exploit cross-selling potentials. One of the recent additions to Amazon.com is the “Great Buy” feature. It offers the customer the current product together with another product in a bundle at a special price (Figure 1b). The two products must have a complementary relationship to be valuable to the customer.

Outcome and Lessons Learned

Amazon.com has clearly utilized a very broad range of different personalization methods. As this site has developed constantly in product variety, functionality, and comfort, it is nearly at the state of the art in this area. The goals stated above are nearly completely fulfilled. The site is easy to navigate, the products are easy to find, and the accuracy of the recommendations seems to be very high, which animates the customer to buy further products.

The “Great Buy” feature is certainly one of the best ways to take advantage of the cross-selling potential, whereas the list of products other customers bought is more useful for the discovery of new, interesting items and navigation. The latter seems to be less accurate compared to the earlier days of the site. This might have something to do with the enormous number of customer profiles that do not provide enough distinctive attributes to form useful clusters.

Ratings and reviews can be considered as a helpful feature; however, there are a number of relatively unqualified comments. To improve the situation, the rate-the-rating feature was introduced (far after the review function itself). While this highlights the more valuable reviews, there is still room for improvement.

APPLICATION AND IMPACT OF PERSONALIZATION: YAHOO!

Yahoo! was one of the first search portals on the Web and one of the first Web sites that applied personalization on a larger scale (Manber, Patel, & Robinson, 2000). In 1996, the My Yahoo! service was introduced. It allows setting up a personal
version of Yahoo! for every user. Not only the content, but also the layout and the appearance of the page can be modified. Yahoo! is a completely adaptable system; therefore, all personalization is based on the data the user entered beforehand. The ZIP code is especially central as a lot of personalized features rely on it. The intelligence of the system lies in the ability to use this data in different situations to tailor the presentation specifically to the user.

**Goal**

The goal of Yahoo! is to bind its users by differentiating from other Web catalogues and search engines, and to provide a fully customizable and integrated portal. As the customer structure of Yahoo! is very heterogeneous, it is a good idea to offer personalization and let users construct an individual start page. Yahoo!’s service is free for
its users; money is mainly earned with advertising and revenue provisions of shopping partners. Thus, the second goal is to make advertising as effective as possible. This can be achieved by selecting banner ads that are likely to be of interest for the user.

### Personalization Techniques

Yahoo! offers an adaptable system that requires the user to explicitly provide information for personalization. The user profile is kept on a server between different visits, thus Yahoo! offers persistent personalization.

My Yahoo! enables registered users to build their own Yahoo! pages. The content is selected as so-called modules. Among the available modules are ones for weather, news, sports results, stock quotes, horoscopes, movie reviews, personal news filters, and many more. Further configuration can be done within these modules. In the example shown in Figure 2, the headline module (a) is set up to show world news from British news agencies and German Formula 1 news. The weather module (b) displays the current weather situation of selected cities only. Modules can be edited or deleted (c) directly on the page. Some of the modules offer an individual default setting that is based on the user’s area.

Not only the content, but also the layout is customizable. The chosen modules can be distributed on multiple pages that in turn consist of two or three columns where modules can be ordered arbitrarily. There are also options like colour sets and themes to change the appearance of My Yahoo!.

**Figure 2. A customized version of the My Yahoo! start page**
Outcome and Lessons Learned

The goal of flexible customization to provide an integrated portal can be considered to be fulfilled. Whether the effectiveness of advertising is increased by personalization cannot be decided with certainty since most of the literature does not mention it. However, as it is vital for Yahoo! to have effective advertising, it can be assumed that it incorporates a relatively advanced system for selecting banner advertisements on an adaptive basis.

Apart from these sophisticated personalization features, it is also essential to design high-quality default pages for people who do not want to customize at all. It turned out that only a small number of users actually customize Yahoo!: most of them take what is offered. The reasons for this may be either that the personalization tools are too complicated, or the users do not need complex personalization as the default page is satisfactory for them. Addressing all types of users also includes not requiring users to enter any personal data and not forcing them to use personalization features.

Yahoo! has decided not to use adaptive methods. They believe that these methods are still not good enough for such a complex site as a user’s behaviour is not sufficiently predictable. People must not be unsure of how the systems work; otherwise, it prevents them from experimenting as they fear breaking something. The people from Yahoo! reckon that any kind of personalization should encourage the users to experiment.

While it is beyond the scope of this article to look at the numerous other cases of personalization, it is worthwhile to mention the large-scale personalization project at ibm.com. Karat, Brodie, Karat, Vergo, and Alpert (2003) describe this in detail, and they provide an overview of the many different personalization techniques used and the reasons for selecting them for the project.

CONCLUSION

Current State of Development

A number of different approaches to personalization are currently used in various applications. Highly sophisticated e-commerce sites usually employ a combination of multiple methods and contain adaptive as well as adaptable elements in a coherent user interface. As the case study of Amazon.com has shown, organisations can add real value for the users by making it more convenient to tailor the presentation to individual needs.

Adaptive methods for personalization are very powerful means to manage information overload and simplify user interfaces. Despite their sophisticated algorithms, they can produce unpredictable and unwanted results. This is an area where further research is necessary as adaptable methods have the serious disadvantage of their configuration remaining static until the user changes it explicitly. This can be a tedious task for most users, which lowers the value even of the most flexible customization services. Hence, there are only a small number of users who actually want to customize their pages.

Directions for Future Research and Development

Improve Group-Based Personalization

The personalization based on a user’s preferences is increasingly being extended by collaborative methods. Instead of recommending items based on the profile of a single user, the system should try to take advantage of other users’ ratings and preferences as well. While this might in fact increase the accuracy of predictions, it raises the issue of proper user selection.
**Personalization Techniques and Their Application**

**New Input Methods**

Methods that allow formulating queries in a natural language rather than in special inhuman query syntax could make the interaction between the user and the system even more personalized and individualistic (Zadrozny, Budzikowska, Chai, Kambhatla, Levesque, & Nicolov, 2000).

**Combination of Explicit and Implicit Data**

Both adaptive and adaptable methods have their strengths and weaknesses, and may be more or less applicable for a particular situation. However, it seems to turn out that the combination of explicit and implicit user data provides the best results. On one hand, the effort of manual customization is minimized, and on the other hand, an adaptive method will not cause much unpredicted results when it is limited by explicit statements. Wells and Wolfers (2000) explain how customers who use online banking services need to make some basic statements about their financial goals and situations. After that, adaptive methods are used to offer financial products and services that are tailored to the customer and suitable for his or her particular needs. However, not all techniques applicable to a certain scenario will be successful in practice. Alpert, Karat, Karat, Brodie, and Vergo (2003) show in their study about users’ attitudes toward adaptive systems that users have a strong desire to always be in full control of all interaction. It is therefore important to carefully analyze the potential acceptance barriers of a designed solution before finally deploying it.

**Future Outlook**

Personalization technologies have found their way out of experimental systems of researchers into commercial applications. They are a powerful means to handle information overload, to make complex information systems more usable for a heterogeneous group of people, and to help online businesses establish personal relations with their customers (one-to-one marketing). Although we have focused on applications of personalization to Web sites, they can be used in a wide range of human-computer interactions. Personalization techniques are the key to mass customization and provide people with a much more individual experience instead of standardized services.

**REFERENCES**


Adaptable Personalization Systems: Systems that can be customized by the user in an explicit manner; that is, the user can change the content, layout, appearance, and so forth to his or her needs.

Adaptive Personalization Systems: These change the presentation implicitly by using secondary data. This data can be obtained from a variety of sources, for example, from the user’s actions, from the behaviour of other users on that site, or based on the currently displayed content.

Decision-Support Systems and Tools: In a wider sense, it can be defined as systems and tools that affect the way people make decisions.

Mass Customization: The customization of products and services for individual customers, but at a mass-production price.

Personalization: An approach to increase the usability of complex information systems and present the user with a comprehensible interface that is tailored to his or her needs and interests.

Recommender Systems: A special type of decision-support system that gives recommendations for further actions or related items.

Tailoring: In the context of personalization, it can be with respect to content or navigation. Content refers to the information being displayed, and navigation refers to the structure of the links that allow the user to move from one page to another.
Chapter 1.14
User-Centered Evaluation of Personalized Web Sites: What’s Unique?

Sherman R. Alpert
IBM T. J. Watson Research Center, USA

John G. Vergo
IBM T. J. Watson Research Center, USA

ABSTRACT
In addition to traditional usability issues, evaluation studies for personalized Web sites and applications must consider concerns specific to these systems. In the general case, usability studies for computer-based applications attempt to determine whether the software, in actual use, meets users’ needs; whether users can accomplish their goals in using the software; whether users can understand and use the application (whether they comprehend what they can do and how); the rate, frequency, and severity of user errors; the rate of and time duration for task completion; and so on. But in the case of user-centered evaluations of personalized Web sites, there are additional questions and issues that must be addressed. In this paper, we present some of these, based on our experience in usability studies of a personalized e-commerce site.

INTRODUCTION
Personalized Web sites attempt to adapt and tailor the user experience to a particular user’s preferences, needs, goals, interests, knowledge, or interaction history. A personalized site adapts its content, content structure, the presentation of information, the inclusion of hyperlinks, or the availability of functionality to each individual user’s characteristics and/or usage behavior. Such a site may place specific information, which it “thinks” you will be interested in, at a distinguished or obvious location on a Web page. Another personalized site may choose to add or
User-Centered Evaluation of Personalized Web Sites

delete specific content or hyperlinks to additional information based on what it “knows” about the current user’s knowledge or interests. An e-commerce site that knows what model laptop you own may only show accessories compatible with that model. A site that displays information about movies and theater schedules may use knowledge of the user’s postal code to display only theaters within n miles of the user’s location. A personalized news site may elect to show (or not) today’s baseball scores, depending on whether the user has viewed this sort of information in previous site visits. A book seller may use knowledge of the books you have ordered in the past to recommend new works by the same author or other authors of the same genre, or may suggest additional books purchased by other users that have bought the same book as you are now ordering. Data about the user, used to drive the site’s personalizations, may be obtained by information explicitly provided by the user and by inferences made by the system based on previous user interactions.

The personalization approach begs many questions. Do personalized Web sites actually improve the user’s experience when using such sites? Do specific personalization features improve and others detract from user experience? Does personalization actually add value to users? Is the site not only usable but acceptable, attractive, and desirable to users?

Personalized Web sites are a specific example of the more general field of adaptive systems. The literature of the evaluation of adaptive systems is replete with evaluative studies of how well the “system” works. These evaluations have focused on algorithms and user model representations for programmatically “implementing” the systems’ adaptive behavior, including determining how well the detection and gathering of implicit information about users’ functions, how appropriately are inferences drawn about users, and how robust are the systems’ techniques for using such information to provide some type of adaptive functionality. For example, evaluations of adaptive systems might consider whether the system’s inferences about the user indeed coincide with the user’s prior behavior (Weibelzahl & Weber, 2003). As another example, “evaluators need to check if [the system’s] inferences or the conclusions drawn by the system concerning the user-computer interaction are correct since it is not necessary that there will be a direct one to one mapping between raw data and their semantically meaningful counterparts” (Gupta & Grover, 2004). Thus many adaptive system evaluations focus on how well the system functions in an (objective) application-centered sense. Many such studies focus on an individual personalization technique, such as recommender systems or collaborative filtering (e.g., Mobasher, Dai, Luo, & Nakagawa, 2001; Zhu & Greiner, 2005). Still others have focused on success of a personalized site as measured by the number of site visits and return visits, number of purchases on an e-commerce site, click-throughs to suggested content, and so forth.

Of course, many of these measures are useful and must be considered in the evaluation of a personalized site. However, evaluations of personalized Web sites must also consider the more subjective user-centered perspective, and the literature is considerably sparser in this regard. User satisfaction is only partially determined by the accuracy of the algorithmic implementation and, further, user satisfaction may not be achieved even in systems that do provide accurate information (Swearingen & Sinha, 2001). In a user-centered design approach, design decisions are based on the experimentally validated value to users of a system’s features and facilities (Vredenburg, Isensee, & Righi, 2001). Thus user-centered evaluations must involve the testing of the system by (and for) users. Some of the existing evaluation literature has suggested using an evaluator who attempts to take the role of a “typical” user of the system; but we have learned from experience (Alpert, Karat, Karat, Brodie, & Vergo, 2003; Karat, Brodie, Karat, Vergo, & Alpert, 2003)
that testing must involve actual intended users of the system because doing so may elicit results that are unexpected based solely on the system developers’ analysis, and reveals more accurately the real user’s perspective.

In the general case, usability studies for computer-based applications attempt to determine whether the software, in actual use, meets users’ requirements; whether the software performs as it should in supporting users as they attempt to accomplish specific tasks; whether typical users are successful in achieving their goals; whether users can understand and use the application (whether they comprehend what they can do and how); the rate, frequency, and severity of user errors; the rate of and time duration for task completion; and so on. The goal is to test “how well the functionality fits user needs, how well the flow through the application fits user tasks, and how well the response of the application fits user expectations” (Usability First, n.d.). Usability testing may involve such abstract issues as the appropriateness of an overall system metaphor and such low level details as the placement of widgets on the screen, the use of color, and the wording of textual content. Usability testing informs the design of interactive software by obtaining the user’s perspective regarding the design, rather than simply relying on the intuitions of the designers and implementers of the software.

Even when researchers have spoken directly to the idea of evaluating the usability of adaptive or personalized systems, they have not addressed the entire problem: Such discussions often revolve around traditional usability issues or around Boolean comparisons of a whole system with and without its entire adaptive functionality (Weibelzahl, Lippitsch, & Weber, 2002). Of course, standard usability issues must be addressed when evaluating adaptive systems: We do care whether users can accomplish their goals using the system, and further, whether they can accomplish their goals faster or more accurately using a system that adapts to their needs than when using a system that does not. And usability evaluations and system evaluations will have interdependencies: If the system’s personalization mechanisms simply do not function well, the user will lose confidence in the system and all of its personalized recommendations (McLaughlin & Herlocke, 2004). But there are other factors that must be considered, issues that go to the heart of the user experience.

As in the design of any software, developers of personalized solutions begin with what they consider to be “good ideas” regarding the sorts of adaptations that would serve users, using intuitions about users and the software’s functionality. And as in case of other interactive software, these intuitions must again be verified by contact with actual users at some time before application deployment. In the case of personalized applications, however, additional questions that go beyond traditional usability must be addressed by user-centered evaluative studies. The functionality offered by personalized applications must not only match users’ needs and, perhaps even more importantly than before, their expectations, but also their desires and level of trust in computational systems. User studies in this domain must not only assess ostensibly quantitative measures such as time-to-task completion, but qualitative issues such as users’ confidence and belief in the system’s recommendations, personalizations, and other adaptive behaviors. We must determine not only whether the system is usable for performing tasks, but also whether it is acceptable to users and enhances, rather than degrades, the user experience. In this paper, we discuss some of these issues that go beyond the purview of traditional usability evaluations.

The issues discussed in this paper are derived from user studies of personalized Web sites in a particular e-commerce context, specifically the Web site of a large computer hardware, software, and services provider (Alpert et al., 2003; Karat et al., 2003). As such, the issues themselves are derived from empirical evaluations. They may be
considered a step moving toward the full realization of the issues and factors that must be addressed in personalized Web site evaluations.

USER-CENTERED PERSONALIZATION MEASURES

As mentioned previously, there are many proposals in the adaptive systems literature aimed at evaluating whether the application or Web site in question performs its adaptations correctly or accurately. And usability studies of (nonpersonalized) Web sites have shown that ease of use can increase the number of revisits and purchases (e.g., Nielsen, 2003). We touch only gently on general usability issues here. Instead, this discussion focuses on the user’s views and opinions of personalized adaptations, not only whether they work as intended, but even if they do so, whether users want the Web site to be making and using inferences and data about the user to influence or direct an adaptive presentation to the user.

Evaluations of adaptive and personalized applications must ultimately address the question, do the adaptive features actually improve the user’s experience when using the site? (see also Chin, 2001). Designers, researchers, and developers may have many ideas for personalized functionality for a Web site that they think would provide users with some benefit. But actual users when confronted with such features may find them useless or, worse, objectionable. The intuitions of the builders of (personalized and all) interactive software must be confirmed by actual potential users of that software.

Here we introduce some of the questions and issues that must be addressed when performing user-centered evaluations of personalized Web sites. These must be addressed in addition to traditional usability concerns, which we will not discuss but that, of course, must be incorporated into the user-centered evaluation. For example, fundamental to any evaluation is whether the site (and its adaptive behaviors) supports users in accomplishing their goals. Or, even more specifically related to personalization, does the inclusion of personalized features make the user more “efficient” and decrease time-to-goal completion. These are important, and ignoring such issues would be foolish. In this paper, we only touch on these more traditional concerns but go further in discussing issues that are of a more subjective nature and relate to the overall user experience, not simply quantitative efficiency measures such as time-to-task completion.

Do Users Want the Personalized Behavior?

Beyond whether site adaptations support the user in accomplishing tasks and goals, the next question we must ask in a user-centered analysis of a personalized site is whether users actually desire the personalizations the site intends or purports to provide. For example, Alpert et al. (2003) found many users were not pleased with a site’s attempts to infer their needs, goals, or interests to thereby provide personalized interactions may be based on specious inferences. “Users say things like ‘don’t stereotype me -- just give me the options because I prefer choosing for myself rather than having the computer tell me what’s good for me’” (Nielsen, 1998).

One type of adaptive system behavior is to intervene while a user is working. The intervention might be to offer advice relating to the user’s plan for achieving particular goals; to offer just-in-time instruction related to information currently in view; to offer additional, related information or an alternative or associated product; and so on. In a user-centered view of the system, the question is, is an intervention wanted or timed correctly?
For example, “Clippy” is the assistive agent in Microsoft Office applications (who by default appears in the form of an interactive paper clip), who offers advice on how to accomplish goals while using the application. The question is, should Clippy intervene while users are working, and when? Overwhelmingly, at least anecdotally, users dislike the assistive agent and its interventions, and many users disable the advice-giving assistance. Problems here include not wishing to be interrupted while focusing on actual work. This can be considered the annoyance factor and may be based on the cognitive load of having to interrupt a plan to focus on advice for an alternative plan. Perhaps more important is the fact that the assistant often does a poor job of inferring the user’s goals and plans. Thus, not only is the interruption unwanted, but the advice is not useful anyway! Due to the extensive use of these application products, many users are wary of computer applications’ attempts to infer their needs and simply do not trust them to be smart enough to do a good job at it.

Questions Regarding the Use of Prior User Behavior for Adaptive Site Behavior

In personalized Web sites, a user may attempt to find or explicitly ask for information on a particular topic, and the site uses information explicitly provided by the user and implicitly inferred about the user to find and (perhaps adaptively) display or recommend pointers to the appropriate information. The first question this raises is, what implicit information does the site use in making personalization decisions? Where (sites, pages, topics, products, etc.) has the user browsed before? Which search hitlist items has the user clicked on before? Does the site use information about the user’s apparent current information needs, for example, current task and inferred goals; recent needs; needs some time in the past; general long-term interests; where the user navigated and the topical or product information the user viewed; or information gathered about other users? How does a personalized site adapt to individual users — that is, on what information are the system’s adaptations based?

First, personalized sites typically use information explicitly provided by users about themselves, in either questionnaires or in the course of performing actual tasks (e.g., the user’s mail zip code is obtained when the user makes a purchase; this can later be used to provide information personalized by location, such as nearby movie theater listings). In the case of multi-user systems such as Web sites, personalized systems may also use information about other users, including their behavior on the site (e.g., for collaborative filtering, what did other users purchase?). In addition, intelligent personalized sites use implicit information gathered about the user, based on the user’s previous behavior on the site. This includes information obtained by observing the user’s previous navigations (e.g., what topics, products, genres did the user look at before?) and clickstream data (e.g., what links did the user click?).

An application-centered evaluation of such a site might ask questions about how well the system’s algorithms perform, for example, does the collaborative filtering engine accurately display the most popular related items? (McLaughlin & Herlocke, 2004). A user-centered evaluation of such an application must ask at least two questions. First, does it make sense — in the context of this site — to use prior user behavior to decide what to present to that user? And, do users want such personalized site features and have it based on their own previous performance and/or the previous behavior and decisions of other users?

Personalized content, based on implicit information, such as previous navigation paths, also met with mixed reactions in our previous studies. The notion of basing content on previous navigation was rated positively by participants overall, but some participants were so adamant in their disapproval of this feature, and their opinions were
so clear and passionately articulated, that they must be considered by developers of personalized sites. One participant stated the explicit-implicit problem succinctly: “I like to have more control and less assumptions made.” Other participants expressed skepticism that this feature could be implemented in a way that would make it generally valuable to them. Specifically, participants expressed some level of concern about a site’s ability to do this well enough to be useful. The overarching message was, “adapting content based on past navigation would be a nice thing, but we don’t believe you can do it well, so don’t do it at all.”

Clearly, using a particular user’s previous behavior is not always useful: “I viewed pages about x before, but I have very different goals and needs now; don’t bias what I can view based on what I did before.” Participants mentioned “shifting goals” as a problem that they believe would not be handled well by systems that use inferred goals to guide current and future presentations. They asserted that their needs, even on a single site, change over time, and they do not want adaptive behavior of the site to be based on their obsolete goals. As Nielsen (1998) puts it, “Having the computer personalize the website to the user assumes that the computer can guess the user’s needs. This is difficult to do and even more difficult when you consider that the same person may have different desires at different times.”

Refining this result further, adapting content based on immediate context was met with favor by our participants. For example, personalized recommendations can be based on the content of the Web page that is currently visible. It appears that using the immediate or current context to influence concurrent or immediately ensuing content is seen to be useful, whereas attempting to infer current goals based on navigation or other information from the past was not universally welcome. Overall, participants agreed that there ought to be a logical “limit” regarding how far back into one’s past history the site should look for implicit information about goals and interests, and further that this limit should be the current “session.” That is, the consensus view was that a user’s past history should have a limited life span: It is acceptable for the site to adapt content based on the user’s current task and navigation context and even the user’s context and history since he/she logged on today, but it should not look at past behavior beyond this point to a disjoint time or session in the past.

A further complicating issue relating to the implicit gathering of information about a user’s past interaction history is: Is all navigation behavior and clickstream data even relevant or meaningful, shifting goals notwithstanding? Therefore exactly when should such information not be collected by the site? Participants in our studies did not want data collected implicitly about them — and “remembered” in their profile — during exploratory sessions on a site unless they specifically authorize it. As one participant explained, “Hey, I might just be knocking around the site for a while...it doesn’t mean anything, and you’ll fill up my profile with a lot of junk if you implicitly collect that information. Wait until I know that I’m really after something before you start collecting data about what I’m doing, and let me tell you when that is.”

**Consideration of Context**

Some personalization features are based on the behavior, attitudes, likes and dislikes, navigation, or purchases of other users. For example, many existing e-commerce sites recommend additional “related” products to the user when a purchase transaction is about to be completed: “People who bought this product also bought products X, Y, and Z.” This is a simple form of collaborative filtering or recommender technologies (e.g., Burke, 1999; Resnick & Varian, 1997; Schäfer, Konstan, & Riedl, 2001). When applied in the context of e-commerce, these technologies use the buying behavior of prior customers to attempt to
“predict” which products a new customer may be interested in purchasing. But in our user studies, some participants expressed objections such as, “I am not like other people, I have different needs.” Additionally, this feature might be perceived simply as an intrusive marketing ploy: One of our participants found it to be an “obnoxious” attempt at marketing more products.

Looking a bit deeper, it appears that user attitudes about collaborative filtering may in fact be influenced by a variety of factors including the type of e-commerce site involved, the particular product being purchased, the type of user, and the reason an item is being purchased. Our previous evaluation studies involved a specific class of user, a particular type of e-commerce site, for a particular category of products and services: Users were business managers buying computer equipment for their enterprises, driven by specific business requirements in their actual jobs. What other customers purchased is not interesting or important to these users. The upshot is that our users had equivocal reactions to a feature that is nonetheless in extensive use: Collaborative filtering is clearly viewed by users as a benefit in other e-commerce settings, such as book and music sellers like Amazon.com. The clear conclusion is that individual personalization strategies, such as recommender systems, are not necessarily effective or beneficial across the spectrum of users and activities. Instead, user opinions and desires regarding such technologies may depend on multiple dimensions of the e-commerce scenario in question, such as the type of product being purchased, whether the purchase is for one’s self or for one’s company, and whether subjective parameters such as taste or genre are relevant to the item being purchased.

Confidence in and Understanding of the Site’s Personalizations

As we can already see, usability measures for personalized systems are significantly interconnected; For example, if a user does not trust that the site is capable of making reasonable inferences regarding his/her current goals, the he/she will not want the site’s (perhaps specious) adaptations. From the user’s perspective, then, an important question is, do users believe computer systems can do the proposed adaptations well (enough to be useful)? Are users confident that a site can perform the personalized adaptation without annoying them or wasting their time? As we have noted, in our previous studies we found that the confidence issue is so important that some participants disclosed that they do not want any information that has been collected implicitly by the site to become a part of their persistent personal profile. Instead, they wanted only information explicitly provided by them to be remembered by the site: “I want the stored information to be based on what I told you, not what you think I’ve said.”

Especially if the site is making decisions regarding the customizing of information shown to individual users, are users confident the site is giving them “good” or the “best” information available? “When the user knows that a screen has been adapted to them, it is natural for them to wonder just what has been adapted to them. What might a different user see? Are they missing out on something?” (Kay, 2001). These thoughts may intensify a user’s lack of confidence in an individually customized site. In a survey of Web users, 94% said that being able to trust the information on a site is very important or somewhat important to them as they decide to (re)visit a Web site (80% said very important; 80% also said that it is very important that the site be easy to navigate) (Consumer Reports WebWatch, 2002).

Confidence in the site also relies heavily on being able to make sense of the site and its behaviors. In a user-centered analysis of a personalized site, one ought to ask, does the user understand what is being adapted due to personalization, and why? The issue is one of determinism and comprehension. “To navigate through web-sites (e.g., presentations, online-shops, or learning
courses) the user requires a mental model of the site’s structure” (Weibelzahl & Weber, 2001, p. 74). It may be difficult to form a coherent model of the structure, content, and behavior of a site when these change over time and circumstance (although Weibelzahl & Weber argue that an adaptive site may require building a less complex mental model than a nonadaptive site). A site may be more difficult to understand and learn to use when multiple instances of the same user-system interaction result in different behavior by the system. Does the user understand what is changing and why it is changing? These are questions that must be addressed in a user-centered evaluation of a personalized site. Does the user understand why the site is “adapting” its behavior and output? Is the system predictable: “Will the user know, with a reasonable degree of certainty, what will happen when (s)he takes an action?” (D’Hertefelt, 2000). Can the user understand how the site behaves and is he/she disconcerted by the fact that the content shown in a specific context is not always the same? Is the user experience degraded because the ability of the user to understand and predict what the site will do is diminished? Paramythis, Totter, and Stephanidis (2001) refer to this issue as transparency. Users want to readily be able to make sense of site behavior, to understand a site’s rationale for displaying particular content. When past behavior or other nonobvious implicit information is used to generate content, the users’ confused reaction — “Where did this come from?” — conveys to users the notion that they are not in control. On the other hand, content whose origin is obvious or readily inferable is met with favor, while content that is based on something the user, or other users, did at some temporally distant time is often met with disapproval.

Users seem to have similar expectations of a dialog with a software application or Web site as with other people. That is, it appears that users desire responses from a Web site to follow logical conversational conventions (Grice, 1975; Searle, 1969); for example, they expect responses to be relevant to the ongoing dialog and to be as informative as required but not more so. This result coincides with the experimental work of Reeves and Nass (1999). After conducting numerous studies, Reeves and Nass concluded that users’ expectations regarding their interactions with computers and other media are based in large part on social interactions in real life, that is, users expect computers to obey rules that come from the world of interpersonal interaction. When a site provides more information — or, in some manner, other information — than what is expected, the user may not immediately know what to make of the “extra” information. In discussing many features, we heard from our users that they want to understand, without difficulty, why and how the computer side of the ongoing bilateral conversation, represented by the content the Web site displays, chose to “say” what it has displayed.

Further, does a user’s need to understand what is happening vary with the type of personalized site? If personalized adaptation is based on, say, previous navigation, the resultant application behavior may not always be clear to the user. For example, the simple collaborative filtering mechanism implemented on the Amazon site is explained as “Customers who bought [this book] also bought…” Thus the attempt is made to have users understand this simple personalization. On the other hand, when personalized results for a particular search query result in differing output at different times, does the user understand why the results differ? During usability evaluation, we must also be sure that when a site does explain its behavior, the explanation is satisfactorily informative from the user’s perspective.

Andersen, Andersen, and Hansen (2001) assert that users of adaptive e-commerce sites wish to be surprised by the site, but in a very different manner than what is being discussed here. They mean pleasantly surprised in terms of value-added services as a reward for using the site. First, they state that users wish to see recommended products related to what they are already purchasing. We
saw in our user study that this is dependent on the e-commerce context: Users of book seller sites (as in Andersen et al.) may wish to have this feature, whereas in many other product-purchasing contexts this is not a welcome feature. Andersen et al. also state that users wish to be \textit{surprised} by special offers associated with use of the site (e.g., lower prices than in brick-and-mortar stores and flexible delivery options). On the other hand, users do not wish to be \textit{surprised} in the conventional sense; if they ask, “What color is the sky?” they do not expect a response about shoe sizes in North America.

Confidence and trust in a site of course also touch on the broader issue of privacy. The importance of this issue is brought into relief by a survey in which Web users were asked to rate their confidence in various types of organizations and enterprises (businesses, newspapers, banks, charities, government, etc.). Web sites rated among the lowest, and the overwhelming reason was lack of trust regarding the handling of private information (Consumer Reports WebWatch, 2002). For a given site then, usability evaluations must also determine whether users are willing to divulge information about themselves to the site. Do users of the site under evaluation believe that the value added by the site’s personalization features is worth the user-perceived risk in disclosing private information? Are they confident that the site will use the personal information in ethical ways? Is an official privacy statement — in which the site spells out a legally binding description of how it will use and disseminate (or not) personal information — necessary?

**Who is in Control?**

In our previous studies regarding personalized Web sites, the most important issue to users was their fervent desire to be — or at least feel like they are — “in control.” “The feeling of security experienced by a user of an interactive system is determined by the user’s feeling of control of the interactive system” (D’Hertefelt, 2000). The issue is one of trust, and again has interdependencies with other issues, such as confidence and comprehensibility. “An interactive system that allows the user to feel in control should in the first place be comprehensible” (D’Hertefelt, 2000). But also, does the user trust that the inferences the personalized site has made about him are correct? “People usually do not know the reasons behind some of the personal recommendations, which often results in their distrust of the personalization mechanism” (Wang & Lin, 2003). Is there a way for the user to understand the kinds of inferences the site is making? Should there be? Can the user correct the inferences and other profile information? How granular is the control — that is, is it simply on/off for all adaptation, for subsets of the adaptation features? What is the accuracy rate required for the adaptation to become a net benefit to the user? If content or features are elided as a result of inferences made about the user, can the user easily “see” what they are missing: While users may feel that at times personalization may be useful, they also believe the application must provide ways to view the “full,” nonpersonalized or nonadapted content or information (Nielsen, 1998). Can the user view, and even modify, his/her personalization profile? These are the sorts of questions that must be addressed in a user-centered evaluation of a personalized Web site.

In our studies, this control issue emerged most strikingly in terms of being in control of one’s profile, the persistently stored personal information collected by the site about the user. As noted previously with regard to confidence and trust in a site, users ask, “What are you going to do with this data; can I trust you to maintain ‘intimate’ information about me?” While this result is certainly no surprise, it is prevalent among all users and is a strongly emotional issue. As such, it can be considered a “deal breaker” — users will not provide personal information and will not perform transactions on a Web site that in their opinion cannot be fully trusted to use profile information.
in an ethical and private manner. Andersen et al. (2001) also discuss this issue. Users want to be able to review, modify, delete, or add personal information to their profile at any time. This should be another focus of usability evaluations of personalized Web sites, to determine how the users of the specific site feel regarding these issues and what might be done to ameliorate their concerns (for example, Kay and her colleagues have attempted to address these sorts of “feeling in control” issues in personalized systems with scrutable user models and adaptations; see, e.g., Kay, 2001; Czarkowski & Kay, 2005).

Of course, there is a delicate balance that must be maintained between adaptability of the system and the user’s sense of control: “Increasing the flexibility can also increase complexity and diminish comprehensibility” (D’Hertefelt, 2000). Thus the need for actually testing personalized sites with users is further intensified — personalized systems attempt to balance adaptability versus users’ sense of control, and usability testing must be performed to determine whether the site has succeeded in this balancing act.

**How Much Work is Involved on the Part of the User?**

Web site users are loath to spend extra time and effort to enter personal information, even to enable a site to “know” more about them for personalization purposes. Nielsen (1998) invokes the notion of the paradox of the active user to explain this user behavior and preference. When encountering a new application, users wish to immediately engage in actual tasks, tasks that further their own goals. This is especially true when visiting Web sites; users have a specific goal in mind (e.g., purchasing a CD) when they visit an e-commerce site, and have little desire to first enter into an extended session of entering information about themselves before being able to accomplish that goal. “Web users are extremely impatient and want to get something useful out of a site immediately” (Nielsen, 1998).

An alternative to lengthy data-entering sessions is a technique known as “permission marketing” (Godin, 1999), in which a user’s profile or user model is incrementally constructed over time as the user interacts with a service or product provider, such as an e-commerce site. In this scenario, customers are asked only for information sufficient to complete a specific transaction or obtain a particular service, but over time a more complete user profile is obtained. An important aspect of permission marketing is that users receive an observable and immediate benefit in return for the small amount of information explicitly provided to the site and are therefore motivated to comply.

Notably, in our e-commerce studies in which permission marketing was evaluated, users rated the feature very highly. Users also expect that information explicitly entered once will persist — that is, users do not want to be asked by the site to enter the same information at a later time for a subsequent transaction or interaction.

The clear implications for usability evaluations are to determine whether users consider the explicit entering of personal information onerous or annoying, and whether there is consensus among users regarding the implied cost-benefit relationship. Do users ultimately find benefit in the site’s personalization features in spite of the amount or level of work involved in enabling the site to personalize or adapt its interactions?

**WHAT IS THE SAME?**

The aforementioned discussion focuses on what is different about evaluations of personalized Web sites and systems, that is, what distinguishes evaluations of such sites from other Web sites. It enumerates questions and issues that must be addressed in the assessment of personalized sites.
from a user-centered perspective, including asking not only is the site usable, but does it provide functionality and affordances users actually desire, that provide obvious value to users, and that users will choose to use. The evaluation process must address these many questions posed previously. The pragmatics of evaluating these issues is where things are the same as other thorough and comprehensive user-centered assessments. The techniques and approach are the same, the focus and issues addressed are expanded.

As for other types of Web sites, user-centered evaluations of personalized sites certainly ought to involve a system prototype that users can view and perhaps minimally interact with. Upon deciding what personalized features you wish to potentially incorporate into the site, a prototype that reifies those features should be constructed. Evaluation studies should be performed early in the design process using site prototypes; this can be a money- and effort-saving technique if designers’ initial intuitions regarding adaptive features and functionality do not match users’ opinions of them. This savings may be greater in the case of adaptive functionality than traditional software because the former may involve techniques that are difficult and time consuming to implement. As Paramythis et al. (2001) suggest, “Eliciting user feedback regarding the modeling process requires that at least a prototype of the system exists.” Taking this a step further, initial user testing should begin when only a prototype exists, that is, before the site is fully implemented, so that many design and implementation decisions are made prior to the expense and time of building the full site.

A reasonable method of putting the site and its personalization behaviors in front of users without the expense of actually building the system (including its adaptive functionality, which is often complex to implement) is to use paper prototypes, a technique often used in formative usability evaluations. Here, drawings of potential screens are shown to potential users, providing a sense of what the fully implemented system would look like. The images may be skeletal wire frames of the proposed screens or fully rendered realistic images of actual Web pages. Here, drawings of potential screens are shown to potential users who may express their opinions of particular screens, including their content and affordances. Usability study experimenters may be “in control,” demonstrating the proposed site design to focus groups or individual users by walking the study participants through scenarios of interaction, an approach much like the Design Walk-through technique for usability evaluation (Vredenburg et al., 2001). Or participants may discuss what they would choose to do at the interface, what actions they would take to accomplish some scenario-based goal. Experimenters can then show the appropriate image of the site’s response to the user action — for example, a new Web page or the same page with updated information — and the evaluation process iterates over several scenarios.

Evaluations of design ideas might be also performed with prototype Web pages built using presentation software, such as Microsoft® PowerPoint®, thus using digital images on a screen rather than paper. Realistic screen shots can be created using a graphical drawing and editing tool. Taking this notion a step further, in addition to being able to show users static screen images, PowerPoint possesses the ability to easily add programmatic behaviors to provide true interactivity, without the cost of a full, “real” implementation of the actual system. Users cannot only see what screens might look like in the implemented site, but can push buttons, enter text, and in general “use” the interactive prototype, whose behaviors mimic the ultimate implementation, in hands-on, scenario-based interactions. This technique offers a prototype for user testing of greater fidelity than merely showing static screen images and gives users a truer sense of the look-and-feel of the site.
Of course, this implies that driving the evaluation should be realistic scenarios of use of the Web site. Each scenario should involve a “story” in which the user must accomplish a particular goal by using the Web site. The full corpus of scenarios must include ones that exercise the personalization features the site offers. For example, a scenario of use for a computer-purchasing site might involve the goal of purchasing a memory card that is compatible with a particular computer (Alpert et al., 2003; Karat et al., 2003).

One approach might be to contrast performing a particular scenario using two different prototype sites, with and without specific personalization features that might or might not support the user in accomplishing specific goals. Experimenters can walk users through interaction with mock-site prototypes, or users can use, hands on, a minimally interactive prototype, to accomplish this same scenario goal with the two prototypes. Or the walk-through or interaction may occur with only a single prototype, one that incorporates and demonstrates a particular personalization feature. Then, experimenter-led discussions with focus groups and with individual study participants, written and oral questionnaires, time-to-task completion, number of clicks, and keyboard interactions — those materials and procedures typically incorporated into user evaluations of interactive systems — would also be employed (as in the personalization evaluation studies reported in Alpert et al. (2003) and Karat et al. (2003).

For example, questionnaires might list features such as “The site will conduct constrained searches for accessories and upgrades, searching only among those that are compatible with the products you already own,” and study participants would be asked to rate each feature in terms of its value to the participant using a Likert scale ranging from “Highly valuable” to “Not at all valuable.” Evaluations must interrogate whether the personalization features actually helped the user accomplish his/her goals. Questionnaires might include assertions, including such general statements as, “I believe the system helped me in accomplishing my goals” and “The site did not understand very well what I wanted to do and did not offer information helpful to accomplishing my goals,” and more specific statements such as, “When I wanted to purchase memory compatible with my laptop, I found the site helped me in accomplishing my goal” and “When I wanted to purchase memory compatible with my laptop, it was helpful that the site showed only those accessories that are compatible with my particular computer model.” Assertions might also probe the affective issues that have been discussed; for instance, “I liked when the site tried to understand what I was trying to accomplish,” “I’m comfortable with the Web site trying to figure out what my goals are,” and “The site should not try to infer what I’m trying to accomplish and should let me be completely in control.” These too would require Likert ratings from “I strongly agree” to “I strongly disagree.” Group and individual discussions should be sure to focus as well on issues such as the desirability or acceptability of the site’s attempts to infer users’ needs and goals, and how well the system’s intervention indeed facilitated the success of goal achievement.

**CONCLUSION**

Design of personalization or user-adaptive systems (or any software technology) cannot occur in a vacuum, specifically, it cannot usefully proceed without assessing the value and usefulness to users of the concepts proposed and implemented by researchers and developers. Is the overall user experience enhanced due to the inclusion of personalized features and functionality?

We can see that many of the presented user-centered measures interact and overlap: User trust and confidence in a Web site will plainly be affected by the quality of the system’s inferences regarding what the user wants to see; if a user cannot understand why particular information is
being displayed, he/she will have less confidence in the site; if the user is confused by the site's actions, they will feel less in control of the user-system interaction, and so on. The measures may be evaluated with and by users as individual items or joined to form more complex evaluation issues. In either case, the evaluation of personalized Web sites must consider several features, issues, and questions, including:

- Do users desire the particular personalization adaptations provided by the site?
- Do adaptations that apparently succeed in one application context (e.g., collaborative filtering on Amazon.com) necessarily enhance the user experience in the context of the site being evaluated?
- Is the site using appropriate information for its adaptations (e.g., previous navigation, explicitly entered information, previous click-throughs, etc.)?
- Do users find the site's adaptations predictable and acceptable, or unintelligible and even unnerving?
- Are users confused by seeing information they did not explicitly request?
- Do users have confidence and trust in the site based on experiencing its personalized adaptations?
- Are users satisfied with the amount of work they must perform to allow the site to personalize user interactions?
- Do users have an adequate sense of being in control (even though the site may behave differently at different times and circumstances)?
- Are users willing to relinquish control to the extent required to allow the site to personalize interactions?

Many of the questions, issues, and measures we have raised and studied are qualitative (and at times, emotional) in nature. Assessments of personalization features will of course vary across Web site domains, user types, and individual users. Researchers, usability professionals, and developers need to have a broad sensitivity to the issues we have listed, and will be well-served by investing in appropriate user-centered studies before implementing personalization features on their Web sites.

REFERENCES


This work was previously published in Human Computer Interaction Research in Web Design and Evaluation, edited by P. Zaphiris, pp. 257-272, copyright 2007 by Information Science Publishing (an imprint of IGI Global).
ABSTRACT

Usability is integral to software quality. Software developers increasingly acknowledge the importance of user-centered, Web site development. The value of usability engineering and the role of the usability engineer (UE) are less understood. A common assumption is that the UE’s role is only to be a user advocate. To this role, we add the responsibility of addressing concerns of other stakeholders in Web site design and development. We discuss usability engineering and the processes that it encompasses, such as project planning, requirements definition, user-centered design (UCD) and evaluation/testing within the context of traditional software engineering lifecycles. We define the UE’s role throughout a user-centered, Web site development lifecycle. This lifecycle integrates compatible usability engineering processes into software engineering processes, drawing examples from research and experience.

INTRODUCTION

People use the Web in a variety of ways. Their interaction with the Web can be self-motivated or externally motivated; their proficiency novice or expert; their needs and expectations simple or complex. To engineer a successful and satisfactory user experience with a Web site, we need to understand issues such as why people go to
a Web site; what they expect and intend to accomplish at the site; and everything impacting on their experience.

A Web site is the result of a set of processes, usually iterative, beginning with conceptualization, planning and requirements definition, then going on to design, version production, and testing/evaluation, before culminating in the site launch. For usability engineering to be fully integrated into Web site development, its practices must be fully integrated into software development lifecycles (Addelston & O’Connell, 2004, 2005). Lifecycles are structured frameworks for software development activities. For example, Figure 1 incorporates elements that iterative lifecycles typically include. In practice, the sequence and frequency of activities can vary. Research and experience show that including usability in software engineering lifecycles is critical (Mayhew, 1992).

Developing a Web site is a team effort. Each team member has roles and responsibilities. The roles of the usability engineer (UE) are integral to these processes and to the team implementing them, primarily because the UE promotes a user-centered perspective. Software engineering of a Web site addresses a variety of purposes: building a new site; upgrading, refurbishing, maintaining or introducing new information or functionality to an existing site; and replacing a legacy site. These purposes track to the goals of the Web site providers. Software engineering is not inherently user centered. It becomes user centered by incorporating usability engineering. User-centered processes for Web site development are compatible and simultaneous with software engineering lifecycle processes.

Building usability into a Web site requires user-centered processes. Such processes require defined roles and activities, which, in turn, depend on common definitions of concepts, inputs, outputs, and tools. From the start, team members must share an understanding of users’ attributes and needs. This understanding underpins the collaboration necessary to incorporate user-centered processes into Web site development.

Figure 1. A generic, variable sequence, iterative Web site development lifecycle illustrates points where usability engineering is most beneficial

Note: With the exception of version production, each of the activities in the outer ovals includes both usability engineering and software engineering processes. In practice, the sequence and frequency of activities can vary.
Guided by Figure 1, this chapter addresses the following topics:

- Usability
- Users
- User interface
- Usability engineering
- Software engineering
- Integrating usability engineering into software engineering lifecycles
- Lifecycle activities

The chapter ends with a summary and references.

USABILITY

People outside the field of usability engineering sometimes consider usability to be obvious, but vague and unstructured — something common sense can recognize and accomplish. Sometimes they are surprised to learn that the field has its own definitions and established processes. Although those people are happy to define usability as “I know it when I see it,” for UEs, a strict definition underlies our focus on users’ needs and our goal of meeting those needs through usability engineering processes. This chapter discusses users’ needs in the context of human-computer interaction (HCI), specifically as users interact with Web sites.

The International Organization for Standardization (ISO) defines usability through the attributes of users’ interactions with software products in specific contexts of use: efficiency, effectiveness, and user satisfaction. We “boil these attributes down” to two outcomes: (1) user success and (2) user satisfaction (1998). The ISO definition implies that usable software must be accessible.

Throughout this chapter, we draw illustrations from our work in a specialized branch of usability called accessibility. Accessibility enables people with disabilities to experience success and satisfaction with software to a degree comparable to that enjoyed by people without disabilities. Although some authors treat accessibility as distinct from usability, we consider it to be a subdomain of usability in which the users are people with physical and/or cognitive disabilities (Hix & O’Connell, 2005).

USERS

In the context of this chapter, users are people who interact with Web sites. In the sense in which we use the term, users are also known as end users, the people who visit Web sites and interact with their contents. The term user excludes people employed in a Web site project, for example, the UEs. It excludes the site’s providers and others who have any stake in the Web site. People close to the project can be too technical or too expert in a domain to represent a user who does not have the same training, goals or background. Those close to the project run a high risk of unintentionally clouding their view of users’ needs with their own commitment to achieving the project’s goals.

Many variables are inherent to users (Bias & Karat, 2005). Some are intrinsic, for example, age; gender; experience with technology; intellectual or aesthetic preferences; interaction styles; and the presence or absence of physical or cognitive disabilities. Other variables, such as employer goals and working environment are extrinsic, but affect the user experience. Many user attributes can decline with age, for example, memory and perception of color contrast (O’Connell, in press).

Each user brings a cluster of capabilities and limitations to any interaction with a Web site. These are the well-documented human capabilities and limitations in perception, manual dexterity, memory, problem solving, and decision making (e.g., Baddeley, 1990; Brown & Deffenbacher, 1979; Mayer, 1992). For example, the limitations of working memory are well known: seven plus
or minus two “chunks” (sets of items), with the size of a chunk varying depending on the user’s experience (Miller, 1956). Research has found that working memory can be further limited by environmental characteristics, for example, noise, fatigue, perceived time pressure, and other sources of stress (Bradley, 1990). Correctly applying such research findings to Web site design is the UE’s responsibility.

Some user attributes are physical. The term user includes people with disabilities as well as people without disabilities. Users with physical disabilities require a Web site that accommodates those disabilities. For example, a user with low manual dexterity may be a skilled problem solver, but have difficulty clicking on small screen elements with a mouse. As content density increases on Web pages, the need for small controls rises. Consider a page with a long list of expandable or contractible menu options. By offering the expanded state as the default, the site accommodates users with low manual dexterity, relieving them of the need to click on the tiny controls. Tradeoffs are inevitable. The user who cannot expand or contract the list must contend with content density. Accommodating users’ capabilities and limitations is rarely simple. Sometimes we must accommodate several factors at once. The user with the manual dexterity disability may also have a visual disability brought about by age.

The World Wide Web Consortium (W3C) considers the needs of users with a wide variety of disabilities, the W3C gives guidelines and success criteria for assessing accessibility (W3C, 2004). For users with special needs, an accessible Web site provides perceivable content; operable content components; understandable content and controls; and content compatibility with other technologies (Caldwell, Chisholm, Slatin, & Vanderheiden, 2005).

Publishing a Web site makes it available to the world, but it is not possible to design for the world. Certainly, there are generic commonalities across Web sites. Any Web site can have hyperlinks, controls, and information architecture. But, any Web site will also have unique features to meet the needs of its targeted audience. Designers need to know the characteristics of the site’s targeted audience to promote their success and satisfaction at the site. They must also project future user characteristics. For example, the intended audience needs to include users with disabilities, even if none of the current users has a disability. Today’s able user may have a disability tomorrow. Age brings disabilities (O’Connell, in press). New employees may have disabilities.

A Web site goal is to enable its users to experience success and satisfaction by accommodating each user attribute of the set of targeted users as well as the unique constellation of attributes defining each individual user. However, the “average” user is a myth. Designers need to understand the range of each attribute within a group of users and to recognize that users will fall at all points along the range (e.g., low to high spatial ability).

To help us accommodate these ranges while still focusing on targeted users, we group users into classes, rather than design for them as individuals. We define user classes by people’s participation in a set of user attributes. For example, for a theater-ticket vending site, the senior citizen user class will include the attributes, aged 65 or over and partial memory loss. The first attribute qualifies these users for a discount; the second tells UEs that the interface needs to remind users of the time of the show. Other user classes for this site would include adults, children, and members of theater clubs, each class having its own attributes.

### User Participation in Web Site Design

It is essential to include users in the process of Web site design because users will do what they will, not necessarily what the designers or developers want or expect them to do. User behavior can be anticipated up to a point, but not predicted in exact detail. Not being intimately familiar with
The Usability Engineering Behind User-Centered Processes for Web Site Development Lifecycles

The design of a site, users will try to do what seems logical to them, even though their logic may conflict with the designer’s or developer’s logic. According to Shneiderman (1998), users do not make errors of intention because they are doing what seems logical to them in the context of the task they are trying to complete. Users do not try to get lost, but they do get lost.

User involvement in design extends our understanding of human behavior within the context of the site and its stakeholders’ goals. In some projects, users participate in focus groups and interviews or respond to surveys. They join in conceptualization and design discussions. Observing users interact with a Web site provides the most valuable new insight, taking UEs to a level of understanding not otherwise possible. These observations are combined and analyzed to develop profiles of the range of expected interactions. Observing users interacting with the site identifies unanticipated needs and behaviors. Analyzing observational data produces requirements exceeding what programming logic, technical expertise, and the best of intentions can identify.

Some software engineering processes eliminate user participation but purport to accommodate a user orientation. These include assigning a user surrogate and processing user information channeled through a nonuser, such as the user’s manager. In contrast, usability engineering relies on close interaction with real users at the strategic points where their input is most crucial. It documents input from the users who actually participated in lifecycle activities.

Knowledge gained from usability engineering cannot be gained by other means. This is not to say that users drive design. Users are not designers. It is a key role of the UE to translate the understanding of users and their needs into design recommendations. It is a key role of designers and developers to incorporate usability recommendations into interactive interfaces that promote users’ success and satisfaction.

Users’ Mental Models

Along with their capabilities and limitations, users bring with them their previous experience of computers and Web sites. We assume they have built mental models, that is, psychological representations of the ways in which computers and Web sites work (Carroll, 1990; Johnson-Laird, 1983; Van Der Veer & Del Carmen Puerta Melguizo, 2003).

Highly experienced users can have mental models of different categories of Web sites, for example, sites for entertainment, information gathering, and e-commerce. According to the user’s mental model, an entertainment site should use bright colors and animation, whereas an information site should use subdued colors and minimal animation, for example, only to demonstrate an integral concept. Abstracting and representing users’ mental models is another job for the UE. Tools for doing this include cognitive task analysis (e.g., Crandall, Kline, & Hoffman, 2006).

User expectations for Web sites are based on their experience with the Internet. They have expectations for the behavior of controls, for example, the back and forward button, Tab key, and Enter key. They have expectations for hyperlinks, for example, a link once activated will change its color. Some expect to be told whether clicking on a link will take them outside the site. Sometimes, users worry about whether they will be able to get back to where they are if they click on a link because they have been unable to form a definite mental model for link behavior. This uncertainty can arise when they have experienced links behaving inconsistently across sites.

USER INTERFACE

In one sense, a user interface (UI) is software that people use to interact with technology. For UEs it is a matter of layers. Above the underlying
technology is the important look-and-feel layer (Garrett, 2002). Feel refers to more than the point and click paradigm; it also refers to what users hear, for example, a sound associated with an alarm. The UI includes both the user’s actual and expected mode of interaction with a Web site, for example, keyboard, mouse, or speech. Whether the designers intend it or not, because of mental models the UI includes implied functionality based on similarities to familiar software. In the broadest sense, the UI is the virtual place where the user’s mental model, meets the designers’ system model (Bolt, 1984). Aligning these models is a goal of usability engineering (Norman & Draper, 1986).

**USABILITY ENGINEERING**

Usability engineering is a set of defined, user-centered processes, grounded in research and experience-based principles. The purpose of usability engineering in Web development is to raise the potential for users’ success and satisfaction and, thereby, to support Web site providers’ goals. The UE must understand the complex set of variables residing in any user group and apply this understanding to promote users’ success and satisfaction. This understanding is what makes usability engineering critical to achieving Web site usability. Because people’s styles of interacting with technology change as technology progresses, usability engineering is a continually evolving field informed by applied research in human interaction with technology.

The UE applies expertise not usually found in other software development team members to make an essential contribution to the quality of a Web site. As noted by Bias and Karat, “good usability is not standard for most Web sites…” (2005, p. 2). When usability engineering is not part of Web site development, the team faces a high risk that, at the least, the site will not promote users’ success and satisfaction; at worst, it will inhibit users’ success and satisfaction and thereby prevent achievement of the site providers’ goals. This outcome would be a disservice to the providers and other stakeholders.

Usability engineering is practiced by UEs who typically have training and experience in a variety of disciplines. In the case of Web work, relevant experience is in fields such as psychology, HCI, testing protocols, and design. The UEs’ work is distinguished by its orientation toward the user, but usability engineering is more than user advocacy. For example, in addition to understanding users’ mental models and expectations, the UE must also be well versed in technology, standards, and laws. To create feasible design approaches, the UE must understand organizational and project goals. To integrate usability into the lifecycle, the UE must be able to communicate with users, stakeholders, and other members of the Web site development team.

A Web site has several objectives. While delivering information and functionality, it also bears the responsibilities of putting the site providers’ best foot forward and achieving their business goals. A Web site’s usability influences users’ impressions of the site providers’ integrity and trustworthiness. UEs are qualified to coordinate the factors necessary to meet the needs of both users and site providers. As we show later, the means to this end is to incorporate usability throughout a software development lifecycle.

**Usability Engineering Principles**

Usability engineering is a multi-disciplinary field. With roots in human factors, it also draws on disciplines such as software engineering, linguistics, biology, cognitive psychology, technology, and graphic design. These fields’ diverse contributions are documented in usability principles, a set of research and experience-based, widely accepted guidelines for achieving usability (e.g., Koyani et al., 2003; Mayhew, 1992; Shneiderman & Plaisant, 2004). UEs continually update usability principles
Usability principles empower design strategies to meet users’ needs and expectations, while avoiding unnecessary, inappropriate features that burden schedules and budgets.

Most usability principles trace to human capabilities and limitations. A user with color deficiencies can detect changes in brightness or shape but cannot distinguish between certain colors. Therefore one usability principle tells us never to rely on color to convey meaning. An important usability principle requires taking steps to assure users always feel in control. Progress bars and hourglasses give users a sense of where they are in a process, contributing to their sense of control. Another principle obliges UEs to direct users through design, for example, leading them through a Web-based form, by following their expectations, for example, whether the family name should precede or follow the last name.

Stakeholders

Everyone who has any interest in a Web site is a stakeholder, whether they are company officers, the marketing team, or the system administrators in charge of maintaining the site. Site providers are the stakeholders who finance the Web site and set the business goals. They stipulate the purpose of the site from the business point of view. They allocate resources and set policy.

Although some consider UEs to be principally user advocates, UEs have another equally important job. That is to address the concerns of stakeholders during software development. It is the UE’s responsibility to make sure that the design, while promoting users’ success and satisfaction, also promotes the aims of the Web site’s stakeholders.

Sometimes stakeholders have misconceptions about the worth of usability engineering or how it integrates with software engineering. Recognizing and addressing this fact is an essential task for UEs; they must remedy these misconceptions or expect little chance of successfully integrating usability engineering in a Web development project (Hix & O’Connell, 2005). For example, it is not uncommon for stakeholders to mistakenly consider usability engineering an expensive add-on that puts users’ needs above all others. To counter this misconception, the UE will give examples of usability as a cost-effective and cost-justifiable, integral contributor to achieving the site providers’ goals while instilling usability (e.g., Mayhew & Tremaine, 2005; Web Accessibility Initiative, n.d.). Another means to overcoming stakeholder misconceptions is to bring them in to watch early usability evaluation sessions where they see users encountering usability problems. When they return for later sessions after changes have been made to the UI, stakeholders observe users experiencing the benefits of designing for usability.

SOFTWARE ENGINEERING

Software engineering is the “application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software” (IEEE, 1990, p. 70). A software engineer plays two roles, computer scientist and project manager. On a large Web project, software engineers typically oversee both the technical and management aspects of development. They accomplish this by following lifecycles that stipulate, for example, a project’s phases, methods, activities, inputs, outputs, milestones, documentation, and risk mitigation strategies.

The software engineering view of usability has not always coincided with the UE’s definition or outlook. Historically, software engineering’s attention to usability was largely confined to summative testing activities (Hix & Hartson, 1993), validation at project’s end that people can use a Web site. Now, software engineers more often consider usability a valuable part of Web site development lifecycles.
INTEGRATING USABILITY ENGINEERING INTO SOFTWARE ENGINEERING LIFECYCLES

Software engineering lifecycles are hospitable to usability engineering. Commonalities between software engineering and usability engineering facilitate this compatibility. The two professions share tools such as use cases, although they sometimes employ them differently. They have the common goal of delivering quality Web sites, on time and on budget, to satisfied users, customers, and other stakeholders. They share terminology, but sometimes with different meanings or connotations. For example, in software engineering, the word *interface* primarily means a connection between two components of a software system, whereas, to a UE, *interface* first and foremost denotes the human-computer interface.

Software engineering and usability engineering processes can occur in parallel because their activities and outputs are compatible. Sometimes these processes are rigid, but the constraints of developing Web sites in real time against tight schedules and tighter budgets drive a trend toward adaptability. This trend emphasizes the need for a UE. In this fast-paced environment, users on the development team can be few and their involvement infrequent. In such a case, a UE draws on knowledge of the field, for example, usability principles and knowledge of users, to aid in the development of usable Web sites.

LIFECYCLE ACTIVITIES

Usability engineering has corresponding activities for most software engineering activities. Not all lifecycles incorporate the same activities. Activity sequence and frequency can vary. Some activities can be simultaneous. Each activity has goals, inputs, processes, and products.

In Figure 1, we present a high-level view of a user-centered software engineering lifecycle.

We use a generic lifecycle where all activities are connected to each other, feeding output into subsequent activities. The sequence of activities is adaptable. Within each activity, usability engineering and software engineering processes occur simultaneously.

Table 1 sets out some activities and processes of software development lifecycles. In addition to usability in general, the goal or objective of these activities and processes is to provide accessibility for people with disabilities. This table is not comprehensive, but it demonstrates how some usability engineering processes can be integrated into a software development lifecycle.

Although we maintain the importance of usability engineering throughout Web site development (e.g., Murphy, Marquis, Nichols, Kennedy, & Mingay, 2001), our discussion zooms in on the most critical iterative process areas. These core activities are project planning, requirements definition, design, and evaluation/testing. Although these are a subset of all possible software engineering activities, we designate these activities as core because they represent the “make it or break it” points in a typical software engineering lifecycle where usability engineering must be addressed. However, none of these activities stands alone. So, we place them within the context of other typical iterative lifecycle activities such as product conceptualization and version production.

Project Planning

A Web site project starts as an idea. This can be a new idea encapsulating the site providers’ vision. More often, at the end of a lifecycle, a team returns to product conceptualization when they evaluate an existing site and begin to plan future versions. In either case, the team needs a blueprint for the steps between the original concept and the insertion of the final product into the workplace. This blueprint is called the project plan.

Formal software engineering lifecycles start with project planning. Successful project planning
Table 1. Usability engineering activities and processes during a software development lifecycle (Partially based on Addelston & O’Connell, 2005; Mayhew, 1992)

<table>
<thead>
<tr>
<th>Software &amp; Usability Engineering Lifecycle Activities</th>
<th>Examples of Usability Engineering Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Product Conceptualization</strong></td>
<td>Literature review, for example, to check latest research results</td>
</tr>
<tr>
<td></td>
<td>Expert review of competition or legacy Web site</td>
</tr>
<tr>
<td></td>
<td>Brainstorming</td>
</tr>
<tr>
<td></td>
<td>Collection and analysis of users’ feedback on legacy Web site</td>
</tr>
<tr>
<td><strong>Project Planning</strong></td>
<td>Overview planning, for example, project summary with requirements</td>
</tr>
<tr>
<td></td>
<td>Project organization planning, for example, identifying and assigning resources, roles, and responsibilities</td>
</tr>
<tr>
<td></td>
<td>Usability risk management planning</td>
</tr>
<tr>
<td></td>
<td>Technical process planning, for example, accessibility processes, equipment, and tools</td>
</tr>
<tr>
<td></td>
<td>Test planning</td>
</tr>
<tr>
<td></td>
<td>Incorporating UE responsibilities into plan</td>
</tr>
<tr>
<td><strong>Requirements Definition</strong></td>
<td>Interviews</td>
</tr>
<tr>
<td></td>
<td>Focus groups</td>
</tr>
<tr>
<td></td>
<td>User profiling</td>
</tr>
<tr>
<td></td>
<td>User class definition</td>
</tr>
<tr>
<td></td>
<td>Translating users’ needs into requirements</td>
</tr>
<tr>
<td></td>
<td>Behavioral and cognitive task analysis</td>
</tr>
<tr>
<td></td>
<td>Persona development</td>
</tr>
<tr>
<td></td>
<td>Setting usability goals for evaluation/testing</td>
</tr>
<tr>
<td></td>
<td>Post-release user surveys</td>
</tr>
<tr>
<td></td>
<td>Analyzing usage statistics</td>
</tr>
<tr>
<td></td>
<td>Contextual user observations</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td>User-centered design</td>
</tr>
<tr>
<td></td>
<td>Interaction design (e.g., use cases)</td>
</tr>
<tr>
<td></td>
<td>Participatory design</td>
</tr>
<tr>
<td></td>
<td>Style guide writing</td>
</tr>
<tr>
<td></td>
<td>Story boarding</td>
</tr>
<tr>
<td></td>
<td>Card sorting</td>
</tr>
<tr>
<td><strong>Version Production</strong></td>
<td>Instantiating design in Web-authoring tools</td>
</tr>
<tr>
<td><strong>Evaluation/Testing (Across versions)</strong></td>
<td>User interface evaluation with user participation (measurement of users’ success and satisfaction)</td>
</tr>
<tr>
<td></td>
<td>Expert review</td>
</tr>
<tr>
<td></td>
<td>Automated testing with accessibility tools</td>
</tr>
<tr>
<td><strong>Launch &amp; Site Maintenance</strong></td>
<td>No usability engineering processes</td>
</tr>
</tbody>
</table>
depends on a clear understanding by all team members of the site providers’ vision as well as an appreciation of the concerns of other stakeholders. Usability engineering and user involvement must be addressed from the start in the plan’s schedule and budget. If not, they are excluded early, with little chance of later inclusion. This is true whether development focuses on a new site or an existing site.

Project planning aims to ensure on-time, on-budget delivery of a Web site that meets its providers’ organizational goals and fulfills requirements. Usability engineering adds the goals of users’ success and satisfaction. Another primary goal is to gain site provider and team acceptance of usability engineering (Hix & O’Connell, 2005). Input to project planning includes output from product conceptualization such as market analyses and everything learned while maintaining a legacy site.

Typical processes involved in project planning are setting schedules; identifying and allocating resources; and stipulating activities, inputs, outputs, and milestones. Usually, planning is the responsibility of software engineers and managers. However, the UE contributes to planning by identifying the points where usability engineering is necessary and identifying the inputs and outputs needed to proceed from one step to the next. For example, the UE will point out the need to design and produce an accessible version with alt tags (alternative text). An alt tag is HTML code associating words with Web site graphics to describe their content. Alt tags must be present before evaluation with a screen reader, an assistive device that reads aloud the contents of Web pages, including alt tags. Screen reader users include people with little or no eyesight. Assistive hardware, such as a screen magnifier, or software devices, such as screen readers, help people with visual disabilities interact with technology. The UE also alerts the team to the need to schedule users with visual disabilities to participate as evaluators.

The product of these processes is a project plan that integrates usability engineering activities with other scheduled project activities and defines all players’ roles, responsibilities, milestones, inputs, and outputs. The unique goals of usability engineering during planning are that the Web site’s design will consider the user and be hospitable to user-centered processes, and that the project will fund, schedule, and provide usability engineering. If these goals are to be achieved, the plan must require usability engineering.

**Requirements Definition**

From the software engineering perspective, requirements definition is a set of processes that identify and document a Web site’s goals in terms of how the site will fulfill its providers’ vision by delivering information and/or functionality. It focuses on user needs assessment. Usability engineering also looks at the site from the users’ perspectives as well to verify that the users’ needs and expectations are being met. It addresses user requirements.

Web sites have functional, system performance, and usability requirements. Functional requirements define what a Web site is supposed to do. For example, a functional requirement for an e-commerce site that sells printers stipulates that the site must display photos of the printers. System performance is a measure of how well the Web site does what it is supposed to do. In our example, a system performance requirement stipulates that the site will deliver a specified photo over a 56K modem in less than two seconds. Usability requirements are sometimes called non-functional requirements. This term is misleading, however, because it diminishes the importance of usability.

Usability is a measure of users’ success and satisfaction with their experience at the site. But usability must also address the goals of the site’s providers. For example, the providers of an e-commerce site benefit from having customers
revisit their site frequently. To achieve this goal, usability requirements stipulate that a user be able to purchase a product, making no more than one error, and give a satisfaction rating of seven on a scale where nine is the maximum positive rating. Users’ success and satisfaction benefit both the users and the site’s providers.

Some interface requirements are yes/no, for example, the Web site will provide anchor links on every page at its top two levels. Others are quantified, for example, users will be able to navigate, with only one click, from a product description page to a page where they can order the product.

Some requirements are standards-driven. Standards can be defined by the site provider, for example, an in-house standard for displaying the company logo. Standards can come from groups such as the ISO, which, for example, gives standards for representing names of languages in its ISO 639 (ISO, 2002).

Sometimes requirements are legally mandated. Take the case of “special needs”, a term that refers to the needs of users with disabilities. The UI of a U.S. federal government site must comply with Section 508 of the Rehabilitation Act of 1973 as amended in 1998. These requirements aim to provide users who have special needs a quality of information accessibility comparable to that of users without disabilities.

The primary goal of requirements definition for the UE is setting usability requirements, the Web site’s obligations to address users’ needs. In usability engineering, users’ needs and Web site requirements are often the same thing.

Ideally, filling requirements will meet users’ needs. However, a Web site can fulfill all functional and system performance requirements, yet still not be usable. In the end, it is the users’ experiences of a Web site that determine whether the site has achieved the site providers’ vision. So, usability engineering promotes a perspective that incorporates the product concept, but expands it with an understanding of targeted users and their needs. Usability engineering brings an understanding of factors that may otherwise not come into the mix, for example, users’ mental models in the context of their capabilities and limitations.

Usability engineering processes during requirements definition start by considering users and their needs within the context of the Web site’s intended purposes. Inputs to this process are the providers’ goals for the site as well as existing information about targeted users and their needs. As other team members set performance and system function requirements, UEs learn whether proposed content will meet not only providers’ goals, but also users’ needs. If the provider’s goal is to inform users about weather conditions, are users people who need to know only tomorrow’s temperature and precipitation forecast, or do they include fishermen interested in tides as well? If the project involves an existing site, UEs address what worked and what did not work for its current users. This understanding empowers UEs to leverage the fact that each user’s experience of the site is a reflection of that user’s needs and expectations.

Requirements definition brings the UE face-to-face with targeted users during interviews, focus groups, and observations in their workplaces, that is, any place where people interact with the Web. The UE develops surveys and studies existing documentation such as usage statistics for a legacy site. However, existing documentation often omits the users’ real-world practices. Therefore, close interaction with users is key. However, the Web introduces the requirement of designing for a wider range of users than a project can usually involve. This factor dictates that the UE rely also on usability principles and knowledge of human capabilities and limitations within the context of Web use. The UE consults resources, such as Web sites on Web usage statistics, to keep abreast of changing interaction behaviors (e.g., Internet World Stats Market Research, 2005; Refsnes Data, 2005). Such resources inform production concep-
tualization as well as requirements definition.

The UE aims for user-informed requirements, but not necessarily user-driven requirements. Although users’ input is integral and respected, the UE must also draw on knowledge of usability engineering to inform recommendations for requirements. For example, when information must be stressed, users may have a preference for many colors and other emphasis techniques on the same display. The UE knows that a user-driven requirement for large blocks of text in upper case, presented in a large variety of strongly contrasting colors will result in adverse effects on user performance (e.g., Koyani et al., 2003). The usability-engineered requirement will specify what portion of the text to emphasize and how to emphasize it in a way that promotes users’ success. For example, the requirement will stipulate initial capitals and no more than three or four compatible colors that can be distinguished by most users with visual color deficiencies.

Once the UE has collected data on users, data analysis occurs. The first step is to define user profiles, descriptions of the Web site’s target population. User profiles record user attributes such as computer literacy; experience in the subject matter and functionality of the site; physical and cognitive capabilities and limitations; special needs; education; mental models; interaction styles; goals at the site; and tasks — all important factors that impact on user interaction with the site. In another process, user class analysis (also called user group analysis), the UE allocates user profiles into groups according to shared attributes, thereby defining user types, such as patients or policy holders.

In behavioral and cognitive task analysis (also called process analysis), the UE develops descriptions of users’ goals and the tasks they perform to accomplish those goals. These definitions are developed within the context of the site providers’ intended purposes for the site. During this key requirements definition activity, the UE studies existing documents and observes users to learn the steps and sequences that they take to accomplish goals. Behavioral analysis documents observable tasks such as receiving information and inputting to the computer. Cognitive task analysis documents users’ mental transformations and decisions. Doing mental arithmetic is an example of a mental transformation. Do users have to keep a running total in mind of their bill on an e-commerce site or does the Web site do it for them? When comparing products as the basis for decision making, is the product information displayed in a way that facilitates comparison, or does the user have to navigate between pages to find comparative data?

Task analyses include steps and work flows. They describe the users’ experience from beginning to end. For example, when users must fill out a form on a secure Web site, the UE documents the experience by identifying the initial steps users need to go through, such as receiving an information package in the mail, locating the URL for the site, locating the user’s unique identifying number for login, all long before the user navigates to the form (Gordon & Gill, 1997; Hackos & Redish, 1998; Kirwan & Ainsworth, 1992; Redish & Wixon, 2003; Schraagen, Chipman, & Shalin, 2000; Vicente, 1999).

During requirements analyses, the UE can develop personas, imaginary representatives of highly specified classes of users (e.g., Head, 2003; Pruitt & Grudin, 2003). Personas reflect what has been learned about users during requirements analysis. They are detailed descriptions of typical users, often given names and illustrated with commercial photographs. Personas become almost real to the team, serving as reminders of typical users’ needs.

The UE tracks the findings of these processes to the site’s intended content to assure that it is presented in a manner that empowers users to achieve their goals at the site. For example, during user class definition, the UE specifies groups of people with disabilities, associating needed assistive devices with the user group. The UE
introduces requirements for the site to present content in a way that makes it accessible via these devices.

In translating user needs into requirements, the UE draws on a variety of sources. The site’s goals are viewed through the lens of users’ needs. For example, consider a case where one of the site’s goals is to enhance the organization’s image by showing its chief technology officer talking about a new technical approach. Considering users with hearing impairments and users with low bandwidths that rule out multimedia, the UE introduces two usability requirements: (1) captions and (2) a text version of the presentation. Meeting these requirements benefits the organization by spreading the message to some who would not otherwise be able to receive it. Thus, the UE addresses the needs of both the users and the organization that provides the Web site.

Usability requirements become goals for later usability evaluation. Setting specific, measurable usability goals with the client provides a quantitative basis for assessing the design against users’ needs (Whiteside, Bennett, & Holtzblatt, 1990).

Although requirements definition produces helpful artifacts such as user profiles, user classes, user task descriptions, and personas, the most important products of requirements definition are the usability requirements that specify the site’s look and feel. Usability requirements become a checklist for everything that must be accomplished to promote successful and satisfactory users’ experiences at the site.

A project benefits from understanding users, the motives behind their actions, and the rationales behind their opinions about the site. When no requirements reflect users and their needs, a project is at high risk of developing an inappropriate site — a site that does not meet its intended users’ needs and expectations. It risks wasting time and resources meeting inappropriate goals. It risks negative impacts on schedule and budget because of the need to retrofit, that is, redesign and recode the UI to correct usability problems.

Although requirements are established at the beginning of a project, they are iteratively reviewed and updated as more is learned about users and the ways that the Web site’s look, feel, functioning, and performance impact users’ success and satisfaction.

**Design**

Incorporating users’ input from requirements definition, the UE participates in developing the site’s information architecture. Information architecture is like a road map; it sets out the paths that users follow to their destinations on a Web site. It is at the heart of design. Impacting more than site navigation, the information architecture impacts a page’s content and layout. The UE’s role is to assure that the information architecture facilitates navigation and makes finding information natural for users.

Important UCD processes, collectively called interaction design, consider the ways that real users attempt to accomplish goals at a Web site. UEs base interaction design on all that they have learned about the users, for example, their age-based capabilities, mental models, and expectations within the context of the goals of the site’s providers. Usability principles provide UEs with rules of thumb that inform UCD decisions. Consider a site intended for senior citizens who expect a prominent link to articles about leisure activities for seniors. The UE considers usability principles on legibility for older users with decreased visual acuity. These principles recommend a large font and a strong contrast between the font and background colors (e.g., Czaja & Lee, 2003).

**User-Centered Design**

Best practices in usability engineering include UCD, a set of usability engineering processes that focus on understanding users, their goals, their strengths and limitations, their work processes — all user attributes that impact how users will
interact with a Web site. The goal of UCD is to achieve users' success and satisfaction by incorporating the users’ perspective into design.

The UE’s multi-disciplinary background adds value to interface design. For example, understanding technology is a prerequisite for designing an accessible Web site. Publishing a Web site is simple nowadays. Adding unneeded features is tempting, just because it is so easy to do. The UE knows how to manage the impact of features, such as animations, on users with disabilities. It is a simple matter to give an information site a bright red background with flashing blue bold titles. The UE understands the biological impacts of such an approach, the potential for the eye to become fatigued because it is unable to focus (Travis, 1991). The UE also knows that animations increase download time and, that therefore, on an informational site, animations can reduce users’ satisfaction.

The UE brings to UCD an understanding of disciplines such as psychology and semiotics, the science of signs and symbols. When incorporating icons into a UI design, for example, it is important to use standard icons to mean what they usually mean and to test any new designs for user comprehension. If a standard icon is used to mean something different from what the users expect, it is likely to cause confusion. Many users will have no idea how to interpret an ambiguous icon in the context of their tasks. With effort, users can learn arbitrary meanings for icons, but they easily forget arbitrary meanings. Icons need text labels to clearly indicate the actions that will occur when they are activated (Horton, 1994).

The UE applies usability principles to participatory design, a UCD process in which users comment on design concepts and perhaps generate their own sketches. Users offer opinions on mock-ups or prototypes. A prototype Web site is like a preview of coming attractions at the cinema. It includes a sampling, but not all of the features and functions of the planned site. Sometimes it is an experiment to investigate UI concepts.

A typical participatory design process is card sorting, where users sort terms that are going to be used in the Web site into groups which they name. The UE combines results from all participants through a statistical technique. Applying usability principles, the UE then derives a meaningful organization of topics for the Web site to inform the information architecture.

We distinguish between user-centered design and inappropriate user-driven design where users’ input translates directly into design directives. Although user-driven design has the admirable quality of being user-focused, it excludes the input of a UE. In turning users’ requests into design decisions without looking at them in light of usability principles, practitioners of user-driven design run a high risk of producing Web sites that, in the end, do not meet users’ needs. Another pitfall of user-driven design is requirements creep that extends schedules and strains budgets as users add inappropriate features and functions that, at worse, will have a negative impact on their experience at the site (Andre & Wickens, 1995).

Ideally, a team of designers, developers, and UEs document UCD decisions in a style guide to promote consistent design. For example, a style guide specifies conventions for screen layout; size, spacing, and location rules for screen elements; and fonts, icons, and color palettes. The style guide evolves, continually updated to record new design decisions.

Design is rarely a one-time effort. Design versions are iteratively evaluated and revised throughout the lifecycle. In many lifecycles, design and version production are interspersed or even simultaneous. In these situations, the UE performs design consultations and evaluates iterative products, informing design decisions with knowledge of users and usability principles. UCD results in a plan for the ways that users will interact with the Web site. The principal product of design is not necessarily a ready-to-publish Web site. It can be a prototype.
Traditionally, during design, the UE has been a user advocate who consults on decisions involving the UI. In this fast-paced age, UEs are more and more involved in design creation and version production, using Web site creation tools.

Although design iterations occur throughout the lifecycle, once part of the site is implemented, programmers are naturally reluctant to make changes. Engineering usability up front reduces the need to request changes after programming. This is why UEs and programmers need to work closely together during UC design.

The principal benefit of having a UE in the loop is that design is more likely to speak to users’ needs. Results include increased productivity, shorter learning times, longer and repeated visits, increased profits, and decreased costs (e.g., Bias & Mayhew, 2005; Kalin, 1999; Mayhew 1999).

Use Cases

A use case is a formal description of ways a product can be used. It consists of a statement of goals with a description of the users and the processes the designers expect them to perform to achieve those goals. Sometimes, a use case is expressed in a sketch. Use cases first come into play in task analysis activities during requirements definition. They are referred to during design.

Use cases provide an example of how a UE can prevent a well-intentioned practice from misrepresenting users. Use cases are the product of a process analysis technique to develop a simple, high-level statement of users’ goals and processes. Use cases are common to the tool kits of both software engineers and usability engineers.

Basing Web design on use cases has strengths and weaknesses. For each module of a system, common processes are written up with the prerequisites for each process, the steps to take for the users and the system, and the changes that will be true after the process is completed. Use cases help to ensure that frequent processes are supported by the system, that they are relatively straightforward, and that the system architecture reflects the process structure.

Use cases, however, do not account for all possible user interactions at the Web site. Use cases tend to stress behavioral tasks, but do not capture cognitive tasks. Use cases do not leave room for users’ unexpected actions at the Web site. Users will naturally do what seems apparent to them, based on the cues given by the UI. A use case does not necessarily represent a natural action for users in the context of the moment. Use cases can put users in the position of having to provide unexpected input to the computer—input that the computer needs but that users do not necessarily know they are supposed to provide.

Use cases make assumptions about the users, for example, that they understand the internal logic of the system the way developers do. Consider novice users who try to use the back button only to discover it does not work because a second browser instance has launched unannounced and without taking focus. The use case depicts users navigating between browser instances, but does not accommodate their expectation to use the back button or the fact that they do not know about the second browser.

The limitations of use cases demonstrate the need for usability engineering. If design relies on use cases but omits a use case for a certain goal/process set, the site will lack important functionality or information. Conversely, use cases not derived from understanding users can result in unnecessary features or information. The UE adds value to use cases by making user-centered recommendations that would not be in the picture otherwise. The UE adds the human dimension to an otherwise limited view of the user as the provider of input to the computer and the recipient of output from the computer. The UE knows that factors such as the user’s thought processes and physical abilities are key to system success.

As with any other tool, use cases must evolve as a project progresses. During updates, the UE introduces the user perspective, incorporating...
The Usability Engineering Behind User-Centered Processes for Web Site Development Lifecycles

what has been learned since the last iteration about how users will interact with the site. Without a UE, the project risks misapplying use cases. For example, avoiding a one-to-one relationship between use case and screen, the UE assures that screens accommodate users’ decision-making strategies and work flows, not someone else’s model of discrete interactions.

Evaluation/Testing

Verification and validation (V&V) are software engineering terms for testing. Verification is iterative testing against requirements. Validation is the final testing against requirements at the end of the lifecycle. Usability evaluation is a set of V&V processes that occurs in conjunction with other V&V activities and is an integral component of an overall V&V approach. In addition to checking for conformance to usability requirements, usability evaluation has the added goal of assessing a wide range of users’ experiences at the site. The UE keeps the door open for new requirements based on the way real users interact with the site. New requirements become input to the next cycle. At project’s end, they become input for new product conceptualization.

Key user-centered, usability evaluation processes entail observing users interacting with a Web site. Activities for formal user observation include writing a test plan; identifying participant users; working with site providers to set usability goals for each user group and task; defining tasks; writing statements of goals that never tell users how to achieve those goals; preparing a user satisfaction survey; preparing ancillary materials such as consent forms; carrying out the observations; analyzing data, and writing a report (Lazar, Murphy, & O’Connell, 2004). These formal processes entail structuring evaluation activities to reflect the tasks identified during requirements definition.

Evaluation draws on the products of all earlier activities. For example, usability goals are based on input from task analysis (Whiteside et al., 1990). Designers also identify features or functions about which they have usability concerns. The UE makes sure that these concerns are addressed in the scenarios to generate information needed to inform UCD decisions.

During a usability evaluation session, users work with scenarios, sometimes behind one-way glass. On the other side of the glass, UEs often employ click-capture software; record numbers and types of user errors; document latency periods when users pause for a significant period of time trying to figure out what to do next; and record critical incidents where users must stop work because of difficulties with the Web site. Ideally, the providers, developers, and stakeholders observe users as they try to accomplish tasks at the site. Observing users interacting with the site can show them the need for changes. Nothing speaks louder about the quality of a Web site than the experiences of its users.

Using a method called think aloud, the UE encourages users to talk about their expectations and reactions while they work with the Web site. The output is metric data on users’ success accompanied by anecdotal data, the users’ own comments on what they were doing and why and what they think of the site. UEs are specially trained to put users at ease during observations and to facilitate the users’ evaluation experience, without telling users how to accomplish their tasks.

After usability observations, users often complete satisfaction surveys on their experience with the Web site. In an effort to foster reliability, the UE only collects and processes feedback on those elements of the site that the user has experienced. These surveys typically collect ratings on a numerical scale to produce metric data. They also offer opportunities for users to elaborate on their experiences.

The UE never relies solely on satisfaction data, but uses it to inform analysis of performance data collected during user interactions with the Web site. Users often report higher levels of
The Usability Engineering Behind User-Centered Processes for Web Site Development Lifecycles

satisfaction than would be expected from their observed performance. This is one reason why UEs interview users about their experience at the end of the session. Another reason is to give users opportunities to bring up points that no one else has anticipated. It is common for developers and other stakeholders to talk with users at this point, too.

Formal usability observations can take place wherever users interact with the Web site. With increasing personal computer sales and the proliferation of mobile devices, people are viewing the same sites from different locations: a conventional office, a crowded cyber café, or a subway train. Each location impacts the user experience differently. A conventional office can have lights that glare on the screen, ringing phones, and frequent interruptions from colleagues. The cyber café can have background noise and poor lighting. The subway can cause breaks in connectivity and the need to view only small chunks of information on a small mobile phone screen. Because of this, when possible, UEs try to hold observations in environments where the site will be used. Otherwise, when possible, they simulate the work environment within the evaluation environment to produce more valid findings.

It is unusual to hold formal usability observations at every development iteration. Indeed, some projects find even one or two rounds of user observations to be cost prohibitive. However, other simple, less expensive processes incorporate users’ perspectives. In an expert review, one or more UEs independently assess an interface against their understanding of users; usability principles; and applicable laws and standards. If more than one UE has performed the expert review, the UEs then meet to discuss and prioritize their findings before discussing them with the rest of the team. Another kind of expert review employs automated accessibility tools, for example, InSight/InFocus (SSB Technologies, 2004). Such tools inspect the Web site code for the UI for conformance with accessibility regulations. They identify violations and recommend remedies.

A lifecycle typically includes several V&V iterations. The product of usability evaluation is a set of recommendations to improve the potential for users’ success and satisfaction with the UI. The principal benefit is an understanding of how users interact with the site (Dumas & Redish, 1999). A unique benefit of usability engineering is the coordination of these recommendations with other stakeholder, organizational, and project goals.

Without a UE, the team risks relying on processes that appear to be usability evaluation, but actually fall short of delivering user-centered products. Examples of pseudo-usability engineering include having stakeholders other than real users provide feedback on interacting with the Web site. Another example is simply asking users to say what they like and dislike. The Web is rife with misinformation about usability engineering. Someone other than a UE trying to engineer usability based on such misinformation can arrive at invalid recommendations that fail to improve the user experience and, at worst, can degrade it. Nothing takes the place of actual success and satisfaction data collected from representative users and interpreted by trained usability engineers.

SUMMARY

We maintain that usability engineering is rigorous, process-based, and addresses needs of stakeholders, such as site providers, as well as users. We have set out a typical software engineering process and discussed key usability engineering contributions. We have demonstrated simultaneous, complementary activities whose products benefit later activities without adversely affecting schedules. We have shown what would be lacking without usability engineering and how the potential of users’ success and satisfaction increases with usability engineering. We stress
that usability engineering is the means to providing successful and satisfactory experiences for Web site users while fulfilling the goals of the site’s providers. The UE’s contribution is integral to Web site development.

REFERENCES


ENDNOTE

1 This material is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

This work was previously published in Human Computer Interaction Research in Web Design and Evaluation, edited by P. Zaphiris, pp. 1-21, copyright 2007 by Information Science Publishing (an imprint of IGI Global).
Chapter 1.16

Personalization Systems and Their Deployment as Web Site Interface Design Decisions

Nanda Kumar
Baruch College, The City University of New York, USA

ABSTRACT

This chapter reviews the different types of personalization systems commonly employed by Web sites and argues that their deployment as Web site interface design decisions may have as big an impact as the personalization systems themselves. To accomplish this, this chapter makes a case for treating Human-Computer Interaction (HCI) issues seriously. It also argues that Web site interface design decisions made by organizations, such as the type and level of personalization employed by a Web site, have a direct impact on the communication capability of that Web site. This chapter also explores the impact of the deployment of personalization systems on users’ loyalty towards the Web site, thus underscoring the practical relevance of these design decisions.

INTRODUCTION

Organizations, in anticipation of the emergence of Web interface as a major point of contact between companies and customers, are beginning to employ a wide variety of technologies to build meaningful relationships with their customers. While Web interface may not be the only point of contact that customers use, organizations are aware of the advantages of using the Web to cater to the needs of the customers. This concept of “self-service” not only reduces costs for the company in the long run, but also increases customer satisfaction by addressing the transactional and the relational needs of the customer. In this chapter, I review the different types of personalization systems commonly employed by Web sites and argue that their deployment as Web site interface design decisions may have as big an impact as the personalization systems themselves. To accom-
plish this, I make a case for treating HCI issues seriously and argue that Web site interface design decisions made by organizations, such as the type and level of personalization employed by a Web site, has a direct impact on the communication capability of that Web site. I also focus on the practical relevance of these Web site design decisions by examining their effect on users’ loyalty towards the Web site.

**BACKGROUND: PERSONALIZATION SYSTEMS**

Most of the technologies and tools that companies use to manage their relationship with their customers usually fall under the banner of Customer Relationship Management (CRM) System. Even though personalization is just one piece of the CRM pie, it is a very crucial piece as effective personalization significantly enhances the ability of the organization to initiate a discourse with its customers to the point where any and all of these dialogues are seamlessly integrated with the database’s historical and transactional information. Based on the data stored in these databases and recent history (the pages customers viewed in the last session), Web sites automatically attempt to improve their organization and presentation of content. These Web sites, armed with a host of appropriate tools — including intelligent agents, recommendation engines and the like — attempt to anticipate the context of the interaction with their customers and personalize each customer’s shopping experience (Andre & Rist, 2002; Billsus, Brunk, Evans, Gladish, & Pazzani, 2002).

Personalization is a process of providing special treatment to a repeat visitor to a Web site by providing relevant information and services based on the visitor’s interests and the context of the interaction (Chiu, 2000; Cingil, Dogac, & Azgin, 2000). Personalization is needed to successfully manage customer relationships, promote the right product the customer is interested in, and

**Figure 1. Overview of personalization techniques**

![Diagram of personalization techniques]

1. **Collection of Visitor Information**
   - Implicit Profiling
   - Explicit Profiling
   - Legacy Data

2. **Personalization Processes**
   - Three Filtering Techniques: Simple, Content Based and Collaborative Filtering
   - Business Rules

3. **Dynamic Assembly and Display of Content as Web Page**
manage content. Most of the advanced personalization might require sophisticated data mining techniques and the ability to display dynamic content without seriously compromising system resources (dynamic display of content will usually mean increased download time).

There are a few well-known techniques for personalization. Rules-based personalization modifies the content of a page based on specific set of business rules. Cross-selling is a classic example of this type of personalization. The key limitation of this technique is that these rules must be specified in advance. Personalization that uses simple filtering techniques determines the content that would be displayed based on predefined groups or classes of visitors and is very similar to personalization based on rules-based techniques. Personalization based on content-based filtering analyzes the “contents of the objects to form a representation of the visitor’s interest” (Chiu, 2000). This would work well for products with a set of key attributes. For example, a Web site can identify the key attributes of movies (VHS, DVD) such as drama, humor, violence, etc., and can recommend movies to its visitors based on similar content. Personalization based on collaborative filtering offers recommendations to a user based on the preferences of like-minded peers. To determine the set of users who have similar tastes, this method collects users’ opinion on a set of products using either explicit or implicit ratings (Chiu, 2000). Please see Figure 1 for an illustration of how a Web site could use all three personalization methods to best serve the customer.

DEPLOYMENT OF PERSONALIZATION SYSTEMS

Profiling

An intelligent way to make the Web site adaptive is to use not only the information provided by the user (such as rating the music and log-in information), but also information that could be collected based on the click-stream trail left behind by the user. These two different sources of collecting information about the consumer are known as explicit and implicit profiling. As the name implies, explicit profiling collects information about a user by directly asking him or her information about himself or herself and product likes and dislikes. This information is collected over a period of time and is stored in the customer database as a profile. Typically, the user would need to log-in in order for the Web site to access the profile and provide personalized content. Even though cookies can be used to store this information on a user’s hard disk, companies prefer to use the log-in approach as this allows the Web site to identify the unique visitor (cookies won’t help if the computer is shared within a family or if the customer accesses the Web site from a different location — say from the office).

Implicit profiling typically tracks the actual behavior of the customer while browsing the Web site. This method of collecting information is transparent to the user. While less intrusive, this method of collecting information has implications for the user’s privacy. Typically, information is collected about the pages the consumer visited, the products he or she looked at and the time that the user spent on these pages. If a (brick and mortar) company has good information systems, the data from explicit and implicit profiling can be merged with the off-line customer information (see legacy user data in Figure 1) to effectively present a seamless Web interface to the customer.

Ideally, a company should use all sources of information it has about the customer. However, when a user visits a shopping Web site (even a repeat user), it would be unsound business practice to expect the user to log-in every time to access personalized content. Hence, a good Web site would use implicit profiling and make a few assumptions about the likes and dislikes of the customer to provide adaptive content to the
customer. For example, if a customer visits a specific product page, it is a good idea to assume that the customer is interested in that particular product and provide content personalized to that user's need. Of course, in most cases, even if the user logs in, the Web site may have little else other than previous purchase history if the user has not provided any specific information on the products he or she likes.

**Personalization as an Interactive Dialogue Between a Web Site and Its Users**

The level and extent of personalization offered by the Web site will have an effect on the communication characteristics of the media. This research argues that different levels of support provided for personalization will specifically impact on the adaptiveness [similar to contingency used by (Burgoon et al., 2000)] of the Web site. This is best illustrated by discussing a real life example using Amazon.com. Appendices 1 to 3 include three screen shots that show the different ways Amazon.com attempts to personalize the experience of the customer. When the user enters the Web site, he or she is invited to log in if desired. Once the user logs in, Appendix 1 shows the Web page that is dynamically created by Amazon.com. This page recommends products to the user based on past purchase history and on the explicit ratings provided by the user to a set of select items. Appendix 2 shows the product page for a book the user is interested in. The column on the left hand side of this page shows the associated related content about the product that is displayed on this page. Appendix 3 shows the page tailor-made for the user based on his recent browsing history and past purchase history. Of course, the scenario described above assumes that the user logged into the Web site at the outset. An intelligent Web site can still adapt its content in its product page by assuming that the user is interested in the product he or she is browsing. Accordingly, the product page shown in screen shot 2 can be personalized even without an explicit log-in by the user.

If the same user were to shop for the book that he is interested in a physical store, he might have approached the sales clerk (or even a friend he had taken along for the shopping trip) for help locating the product. Now, when he mentions to his friend that he is interested in this specific book, music or movie, then it is possible to imagine a conversation happening along the lines discussed above. Of course, the above discourse with the Web site is limited by the need for a shared context. The conversation will not be totally indeterminable in terms of context and content and may not move along in any arbitrary direction as is possible in a conversation with a friend. But, this research argues that there are enough cues in the discourse initiated by the personalization system of Amazon.com that is enough to give the user the impression that the conversation is contingent within that shared context.

**Virtual Communities as Personalized Communities**

To enhance the relationship with the customers, companies can also provide support for virtual communities, as this will facilitate access to free-flowing and unstructured information beyond what is provided by the computer agents (Jones, 1997; Preece, 2001, 2002). For example, companies can aggregate the opinions of consumers on a particular product and present them to a new user who is browsing that product page. Depending on the level of support provided by the Web site, the new user can also get in touch with another consumer he or she might identify with, as is the case with Amazon.com. A recent study (Brown, Tilton, & Woodside, 2002) shows that community features create value for a shopping Web site. Their study showed that community
users accounted for about one-third of the visitors to the e-tailing sites surveyed and that they also generated two-thirds of the sales (2000 transactions worth one million dollars). Practitioners have long argued that having a vibrant community in the form consumer reviews is crucial for the success of e-commerce Web sites such as Amazon.com and Ebay.com (Brown et al., 2002; Kirkpatrick, 2002). Hence providing support for consumer reviews facilitates formation of one type of virtual community and integrating high level of support (user rating and information about the user) for consumer reviews on the product page increases personalization afforded by Web sites as these are relevant comments and opinions by different users presented on the product page.

Deployment of Personalization as Web Site Design Decisions

Reeves, Nass and their colleagues at the Center for the Study of Language and Information at Stanford have shown that even experienced users tend to respond to computers as social entities (Nass, Lombard, Henriksen, & Steur, 1995; Nass, Moon, Fogg, Reeves, & Dryer, 1995; Nass & Steur, 1994). These studies indicate that computer users follow social rules concerning gender stereotypes and politeness, and that these social responses are to the computer as a social entity and not to the programmer. When explicitly asked by the researchers, most users consistently said that social responses to computers were illogical and inappropriate. Yet, under appropriate manipulation, they responded to the computer as though it were a social entity. This, in fact, is the essence of the Theory of Social Response (Moon, 2000; Reeves et al., 1997). Thus I argue that there is value in conceptualizing the Web site as a social actor and that the Web site can be equated to the “agents” mentioned above in terms of source orientation. There are several points-of-contact between a Web site and its users that will result in responses by the users not unlike the way they would respond to a social interaction.

In the light of the above discussions, Web sites should also view deployment of personalization systems as important Web site design decisions that will facilitate or hinder this interactive dialogue between a Web site and its users. In a recent study conducted by this author, the personalization systems deployed by four major Web sites (Amazon.com, BarnesandNoble.com, CDNow.com and Chapters.ca) were compared along with the Web sites’ support for virtual communities. The results of the study showed strong support showing that level of support for personalization systems had an impact on customer loyalty. The results also suggested that Web sites should pay close attention to the way they deploy these personalization systems. Specifically, the results showed that by implementing consumer review support on the product page (as done by Amazon.com), we could simulate support for personalization systems in the absence of personalization systems.

CONCLUSION

In practice, it is advantageous for the Web sites to offer some form of support for personalization or virtual community as this makes the Web site to be perceived as more adaptive. This will facilitate better communication between the Web site and the shoppers, thus leading to higher levels of customer loyalty. Companies do understand that in practical terms it takes a lot more money and effort to acquire a new customer than to keep an existing customer and arguments presented in this chapter throws new light on the role of support for personalization and consumer reviews in increasing customer loyalty.

Good personalization systems can be very expensive to set up. Enabling an e-commerce Web site with the necessary tools to build a vibrant community costs little (especially when compared
to personalization systems), as the community members provide the content. The results of the study offer evidence that Web sites by providing support for consumer reviews could reap the benefits in terms of increased adaptiveness despite offering very little personalization. However, the ways and means of implementing support for personalization and virtual communities deserve further research. Future research should examine the implementation of these features in finer detail. This will help the organizations understand more in depth the trade-offs involved in providing different types and levels of personalization and virtual communities.

REFERENCES


This work was previously published in Web Systems Design and Online Consumer Behavior, edited by Y. Gao, pp. 147-155, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
INTRODUCTION

Traditionally, programming code that is used to construct software user interfaces has been intertwined with the code used to construct the logic of that application’s processing operations (e.g., the business logic involved in transferring funds in a banking application). This tight coupling of user-interface code with processing code has meant that there is a static link between the result of logic operations (e.g., a number produced as the result of an addition operation) and the physical form chosen to present the result of the operation to the user (e.g., how the resulting number is displayed on the screen). This static linkage is, however, not found in instances of natural human-to-human communication.

Humans naturally separate the content and meaning that is to be communicated from how it is to be physically expressed. This creates the ability to choose dynamically the most appropriate encoding system for expressing the content and meaning in the form most suitable for a given situation. This concept of interchangeable physical output can be recreated in software through the use of contemporary design techniques and implementation styles, resulting in interfaces that improve accessibility and usability for the user.

BACKGROUND

This section accordingly reviews certain theories of communication from different disciplines and how they relate to separating the meaning being communicated from the physical form used to convey the meaning.
Claude Shannon (1948), a prominent researcher in the field of communication theory during the 20th century, put forward the idea that meaning is not transmitted in its raw form, but encoded prior to transmission. Although Shannon was primarily working in the field of communication systems and networks such as those used in telephony, his theory has been adopted by those working in the field of human communications. Shannon proposed a five-stage model describing a communication system. Beginning with the first stage of this model, the sender of the communication creates some content and its intended meaning. In the second stage, this content is then encoded into a physical form by the sender and, in the third stage, transmitted to the receiver. Once the communication has been received by the receiver from the sender, it is then at its fourth stage, whereby it is decoded by the receiver. At the fifth and final stage, the content and meaning communicated by the sender become available to the receiver.

An example of how Shannon's (1948) model can be applied to human communication is speech-based communication between two parties. First, the sender of the communication develops some thoughts he or she wishes to transmit to the intended receiver of the communication. Following on from the thought-generation process, the thoughts are then encoded into sound by the vocal cords, and further encoded into a particular language and ontology (i.e., a set of mappings between words and meaning) according to the sender's background. This sound is subsequently transmitted through the air, reaching the receiver's ears where it is decoded by the receiver's auditory system and brain, resulting in the thoughts of the sender finally being available to the receiver.

This split between meaning, its encoding, and the physical transmission of the meaning is recognised in psychology. Psychology considers that there are three stages to receiving data: (a) the receiving of sensory stimuli by a person, (b) the perception of these stimuli into groups and patterns, and (c) the cognitive processing of the meaning with the data (Bruno, 2002). Thus, for example, a receiver may see a shape with four sides (the data) and associate the name square (the meaning) with it. There is accordingly a split between the input a person receives and the meaning he or she cognitively associates with that input.

Consider, for example, the words on this page as an example of the psychological process through which meaning is transmitted. The first stage of the process is where the reader receives sensory stimuli in the form of black and white dots transmitted to the eyes using light waves of varying wavelength. Upon the stimuli reaching the reader, the brain will perceptually group the different dots contained within the received stimuli into shapes and, ultimately, the reader will cognitively associate the names of letters with these shapes and extract the meaning conveyed by the words.

Semiotics, which is the study of signs and their meanings (French, Polovina, Vile, & Park, 2003; Liu, Clarke, Anderson, Stamper, & Abou-Zeid, 2002), also indicates a split between meaning and its physical presentation. Within semiotics, the way something is presented, known as a sign, is considered to be separate from the meaning it conveys. Accordingly, in semiotics there are three main categories of signs: icons, indexes, and symbols. This delineation is, however, not mutually exclusive as a particular sign may contain elements of all categories. Vile and Polovina (2000) define an icon as representative of the physical object it is meant to represent; a symbol as being a set of stimuli, that by agreed convention, have a specific meaning; and indexes as having a direct link to a cause, for example, the change of a mouse pointer from an arrow shape to an hourglass to reflect the busy state of a system.

This classification of the physical representation according to its relationship with the content and meaning it conveys provides further opportunities to distinguish content and meaning...
from its physical presentation, and to classify the different elements of presentation. For example, a shop selling shoes may have a sign outside with a picture of a shoe on it. The image of the shoe is the sign, or the physical presence of the meaning, which in this case is an icon, while the fact that it is a shoe shop is the intended meaning. Equally, this could be represented using the words *shoe shop* as the physical sign, in this case a symbol of the English language, while the meaning is again that of a shoe shop.

This split of content and meaning from its physical presentation, which occurs naturally in human communication, allows for the same content and meaning to be encoded in a variety of different forms and encoding methods. For example, the meaning of “no dogs allowed” can be encoded in a variety of visual images. For instance, there might be (a) an image of a dog with a cross through it, (b) the words “no dogs allowed,” (c) an auditory sequence of sounds forming the words “no dogs allowed,” or (d) the use of tactile alphabets such as Braille, which is used to encode printed writing into a form for the blind. However the content and meaning is conveyed, it remains the same regardless of how it is physically presented.

SOFTWARE ARCHITECTURES FOR CONTENT SEPARATION

For the true separation of presentation from content to occur therefore in software, the content (namely the data or information itself as well as the application’s operations, i.e., its business logic as indicated earlier) is stored in a neutral format. This neutrality is achieved when the content is untainted by presentation considerations. This allows any given content to be translated and displayed in any desired presentation format (e.g., through an HTML [hypertext markup language] Web browser such as Microsoft’s Internet Explorer, as an Adobe Acrobat PDF [Portable Document Format], as an e-book, on a mobile phone, on a personal digital assistant [PDA], or indeed on any other device not mentioned or yet to be invented). The theories of detaching content and meaning from its physical presentation thus give a framework to separate content from presentation. Once that conceptual separation can be made, or at least continually realisable ways toward it are achieved, then this approach can actually be deployed in the design and implementation of computer systems.

There are a number of methods offered by contemporary software languages and architectures to achieve this detachment between the content and meaning, and how the content can thus be displayed. In the sphere of Web development, the extensible markup language (XML) is one such example. XML provides a useful vehicle for separating presentation from content (Quin, 2004a). Essentially, unlike HTML in which the tags are hard coded (e.g., *Head*, *Body*, *H1*, *P*, and so forth), XML allows designers or developers to define their own tags particular to their domain (e.g., *Name*, *Address*, *Account-number*, *Transactions*, *Debits*, *Credits*, and so forth in, say, a banking scenario). How this content is presented has, of course, to be defined by the designer or developer; he or she can no longer rely on the browser to format it by simply recognising the hard-coded HTML tags. The extensible stylesheet language (XSL) is the vehicle to achieve this (Quin, 2004b). Equally, the scaleable vector graphics (SVG) format, based on XML, is another World Wide Web format capable of separating content and meaning from presentation. SVG specifies drawing objects, their dimensions, colour, and so forth, but leaves the determination of presentation modality to the client viewer application (Ferraiolo, Jun, & Jackson, 2003).

Within enterprise systems, this separation can be achieved through the use of object-orientated and *n*-tier design methodologies. Object orientation works through its embodiment of the four goals of software engineering (Booch, 1990; Meyer, 1988; Polovina & Strang, 2004). These
four goals of software engineering, namely (a) abstraction, (b) cohesion, (c) loose coupling, and (d) modularity, determine the principled design of each object that makes up the system. They seek to ensure that the object only performs the functions specific to its role, for example, to display a piece of information or to perform a calculation. Accordingly, these goals seek to ensure that presentation objects only present the information, while logic objects only perform calculations and other business-logic operations. These content objects thus do not concern themselves with how the information is presented to the user; instead these content objects communicate their information via presentation objects to perform this function.

In addition to embodying the four goals of software engineering, object orientation builds on these by providing three further principles: (a) encapsulation, (b) inheritance, and (c) polymorphism (Booch, 1990; Meyer, 1988; Polovina & Strang, 2004). Inheritance allows an object to inherit the characteristics and behaviours of another object. Utilising this feature, it is possible to extend the functionality of an object to include new functionality, which may be new buttons or other interface elements within a user interface. Polymorphism is used to select an object based on its ability to meet a given set of criteria when multiple objects perform similar functions. For example, there may be two objects responsible for displaying the same interface element; both display the same content and meaning, but using different languages. In this scenario, the concept of polymorphism can be used to select the one appropriate for the language native to the user. Thus, object-orientated design can be used to naturally compliment the process of separating content and meaning from its method of presentation.

A common practice within the field of software engineering is to base software designs on common, predefined architectures, referred to as patterns. One pattern, which lends itself well to the separation of content and meaning from its method of presentation, is the n-tier architecture. The n-tier architecture separates the objects used to create the design for a piece of software into layers (Fowler, 2003). The objects contained within each layer perform a specific group of functions, such as data storage. In the three-tier architecture, for example, one layer is responsible for handling the software’s input and output with the user, another handles its business-logic processes, and the final layer handles the persistent storage of information between sessions of the software being executed. Through the use of an n-tier architecture and the separation of the different areas of an application’s design that it creates, it is possible to separate the content from its mode of presentation within software design.

Software engineering’s ability to separate content and meaning from its physical presentation can be aided by some contemporary implementation methods. These methods are based on component architectures that aim to create reusable segments of code that can be executed. This enhances object orientation, which seeks to create reusable segments of software at the source-code level. While there is not much difference in the design, having reusable segments of executable code translates to faster time to change segments, further enhancing the plug-and-play nature of software. Microsoft’s Component Object Model (COM) is a client-side Windows-based component architecture (Microsoft Corporation, 1998). This architecture enables programs to be built as individual components that are linked together using a client application to form a complete software program. This approach to software implementation provides the ability to construct similar pieces of software using the same components, where the functionality is common between the pieces of software. For example, if the storage and logic elements of a piece of software were to remain the same but the user interface were to be changed due to the differing needs of user groups, the same components forming the storage and logic sections could be used for all versions of
Communication + Dynamic Interface = Better User Experience

the software. Furthermore, this could occur while different user components were created to provide the different user interfaces required. This method would reduce the time taken to build and deploy the software amongst a group of diverse users.

Another implementation technique, built around distributed components located on different physical machines, are Web services (MacDonald, 2004). Instead of the components used to build the software being located on the same machine, different components can be placed on different machines. This results in users being able to share and access the same physical instance of objects. This enhances COM, which although it gives access to the same components, forces each user to use different instances of them.

One advantage of Web services is that they allow the existence of different user interfaces while letting users access the same physical objects used for the logic and storage processes. This type of deployment will ensure that all users are accessing the same data through the same logic processes, but allows the flexibility for each user or user group to use an interface that is the most optimal for their needs, be they task- or device-dependant needs.

THE HUMAN-COMPUTER INTERACTION BENEFITS

The human race rarely uses fixed associations between content or meaning and its physical representation. Instead, people encode the meaning into a form appropriate for the situation and purpose of the communication. Communication can be encoded using different ontologies such as different languages and terminology. Communication is thus able to take different physical channels (e.g., sound through the air, or writing on paper), all of which attempt to ensure that the content or meaning is communicated between the parties in the most accurate and efficient manner available for the specific characteristics of the situation. Currently, this is not the case with computer interfaces; contemporary interfaces instead tend to adopt a “one size fits all” approach for the majority of the interface.

In taking this one-size-fits-all approach, content and meaning may not be transmitted to the user in the most accurate form, if it is communicated at all. The characteristics of the situation and participants are not taken into account. This makes the interface harder to use than might be, if it can be used at all. Some users, such as those with a sensory disability or those with a different native language, may not be able to access the information as it has been encoded using an inaccessible physical form (e.g., visual stimuli are inaccessible for the blind). Or it has been encoded using a foreign language, which the user does not understand. This immediately prevents the user from accessing the content and meaning conveyed by that form of presentation.

Equally, terminology can be prohibitive to the ease of use of a user interface. The set of terms that we know the meaning for (i.e., ontology) is based on factors such as the cultural, educational, and social background of the user as well as the geographic area the user inhabits. This leads to different groups of people being familiar with different terminology from those in other groups, although there is some degree of overlap in the ontologies used by the different groups. The user is forced to learn the terminology built into the interface before they can extract the meaning that it conveys. This imposes a learning curve on the user, unless they are already familiar with the particular set of terms used. Hence, by using a one-size-fits-all user interface, some users will find it difficult or impossible to use.

By utilising the facilities offered by contemporary software-engineering practices, it is possible to avoid this one-size-fits-all approach and its inherent disadvantages in terms of human-computer interaction. By allowing the encoding scheme used to present software interfaces to change with different users, interfaces will begin
to mimic the processes used to encode content and meaning that are found in natural human-to-human communication. This change will result in interfaces that are accessible by those who could not previously access them, and will also result in greater ease of use for those who previously had to learn the terminology used within the interface, hence improving interface usability.

**FUTURE TRENDS**

One emerging trend is the use of explicit user modeling to modify the behaviour and presentation of systems based on a user’s historic use of that system (Fischer, 2001). Explicit user modeling involves tracking the preferences and activities of a user over time, and building a model representing that behaviour and associated preferences. This, coupled with the concept of presenting the content and meaning in the form most suitable for the user, holds the ability to tailor the content to a specific individual’s needs. By monitoring how a user receives different types of information over time, a historic pattern can be developed that can subsequently be used to present the content and meaning based on an individual’s actual requirements, not on a generalized set of requirements from a specific group of users.

**CONCLUSION**

Currently, by entwining the association between content and meaning and the physical form used to represent it, software user interfaces do not mimic natural human-to-human communication. Within natural communication, the content and meaning that is to be conveyed is detached from its physical form, and it is only encoded into a physical form at the time of transmission. This timing of the point at which the content and meaning are encoded is important. It gives the flexibility to encode the content and meaning in a form that is suitable for the characteristics of the situation (e.g., the channels available, the languages used by the parties, and the terminology that they know). This ensures that humans communicate with each other in what they consider to be the most appropriate and accurate manner, leading to encoding schemes from which the parties can access the content and meaning in an easy method.

This is not currently the case for software user interfaces, which use a too tightly coupled association between the content and meaning and the physical form used to encode it. By utilising contemporary Web-based or object-orientated component architectures, this problem of fixed encoding schemes can be overcome. Therefore, software user interfaces can more closely mimic natural language encoding and gain all the benefits that it brings.

**REFERENCES**


**KEY TERMS**

**Accessibility:** The measure of whether a person can perform an interaction, access information, or do anything else. It does not measure how well he or she can do it, though.

**Content:** The information, such as thoughts, ideas, and so forth, that someone wishes to communicate. Examples of content could be the ideas and concepts conveyed through this article, the fact that you must stop when a traffic light is red, and so on. Importantly, content is what is to be communicated but not how it is to be communicated.

**Encoding:** Encoding is the process by which the content and meaning that is to be communicated is transformed into a physical form suitable for communication. It involves transforming thoughts and ideas into words, images, actions, and so forth, and then further transforming the words or images into their physical form.

**Object Orientation:** A view of the world based on the notion that it is made up of objects classified by a hierarchical superclass-subclass structure under the most generic superclass (or root) known as an object. For example, a car is a (subclass of) vehicle, a vehicle is a moving object, and a moving object is an object. Hence, a car is an object as the relationship is transitive and, accordingly, a subclass must at least have the attributes and functionality of its superclass(es). Thus, if we provide a generic user-presentation object with a standard interface, then any of its subclasses will conform to that standard interface. This enables the plug and play of any desired subclass according to the user’s encoding and decoding needs.

**Physical Form:** The actual physical means by which thoughts, meaning, concepts, and so forth are conveyed. This, therefore, can take the
Communication + Dynamic Interface = Better User Experience

form of any physical format, such as the writing or displaying of words, the drawing or displaying of images, spoken utterances or other forms of sounds, the carrying out of actions (e.g., bodily gestures), and so forth.

Software Architecture: Rather like the architecture of a building, software architecture describes the principled, structural design of computer software. Contemporary software architectures are multitier (or $n$-tier) in nature. Essentially, these stem from a two-tier architecture in which user-presentation components are separated from the information-content components, hence the two overall tiers. Communication occurs through a standard interface between the tiers. This enables the easy swapping in and out of presentation components, thus enabling information to be encoded into the most appropriate physical form for a given user at any given time.

Usability: A measure of how well someone can use something. Usability, in comparison to accessibility, looks at factors such as ease of use, efficiency, effectiveness, and accuracy. It concentrates on factors of an interaction other than whether someone can perform something, access information, and so forth, which are all handled by accessibility.

This work was previously published in the Encyclopedia of Human Computer Interaction, edited by C. Ghaoui, pp. 85-91, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
ABSTRACT

In this chapter, we introduce the key ideas related to the paradigm of pervasive computing. We discuss its concepts, challenges, and current solutions by dividing it into four research areas. Such division is how we were able to understand what really is involved in pervasive computing at different levels. Our intent is to provide readers with introductory theoretical support in the selected research areas to aid them in their studies of pervasive computing. Within this context, we hope the chapter can be helpful for researchers of pervasive computing, mainly for the beginners, and for students and professors in their academic activities.

INSIDE CHAPTER

The recent advances in hardware and wireless technologies have leveraged the creation of the first experimental pervasive computing scenarios. Due to the belief that these scenarios will be an integral part of future living, research in this field is increasing at a fast pace. Therefore, theoretical and mainly practical studies are of great use as a way of supporting this belief.

Performing such studies, however, implies identifying the intricacies behind pervasive computing. Although its concept is quite simple, understanding these intricacies is a task which scatters across different research fields. Computer
networks, distributed and cognitive systems, software engineering, and user interface design are some of these fields.

Therefore, in this chapter our main objective is to identify and discuss, at an introductory level, some of these intricacies. More specifically, we define four major research areas in pervasive computing, namely pervasive networking, context awareness, pervasive systems development, and pervasive computing middleware. Based on this view, we then take the reader on a journey through the universe of pervasive computing, discussing concepts, challenges, and current solutions.

INTRODUCTION

Today, computing is facing a significant revolution. There is a clear migration from the traditional desktop-based computing to the ubiquitous era, where computing will be spread all around us and seamlessly integrated into our lives. It is this new stage of computing that researchers have named pervasive computing. We can say that it is the accomplishment of the so-called concept of calm technology (Weiser & Brown, 1995), or as Weiser (1993) has said, it “envisions computation primarily in the background where it may not even be noticed” (p. 1). Not surprisingly, these ideas require us to view computers in a totally different way, not only as something we log onto, work on, and log out of when we are finished (Saha & Mukherjee, 2003). Instead, we should see a computer as a portal to a repository of computational resources, making use of them to work on the background and fulfill tasks according to our needs and preferences.

Pervasive computing, also known as ubiquitous computing (Weiser, 1991), has been recognized as the third wave in computer science, following

Figure 1. Sales of mainframes, personal computers, and ubiquitous computing devices
the mainframe and the personal computer ages. Therefore, even if not fully conceived, pervasive computing will be the prevailing paradigm of the 21st century. Observing the graph shown in Figure 1, one can see the sales associated with ubiquitous computing devices follow a fast exponential growth. As more and more facilities, or services, will be available for users of such devices, this growth, even that in a lower rate, will be expected to continue. After all, it is not for nothing that academy and mainly industry are so confident on the pervasive computing paradigm.

Getting a ride on this new trend, the purpose of this chapter is to conduct the reader behind the scenes of pervasive computing, introducing the main concepts and challenges involved in it. The structure of the chapter is illustrated in Figure 2. We start with a first glance at pervasive computing by describing a sample scenario in order to provide the reader with the general concepts. It is presented as an overview of the technological advances that have leveraged the development of pervasive systems, as well as the challenges imposed by pervasive computing scenarios. In the Pervasive Networking section, we present two key concepts for pervasive environments, mobility and host discovery. The notion of context and its importance to pervasive computing will be outlined in the Context in Pervasive Computing section. Next, we present some methods that have been used for developing pervasive systems. More specifically, some techniques that application developers need in order to deal with the inherent characteristics of software for pervasive computing are discussed. Based on this discussion, we then outline in the Middleware for Pervasive Computing section the main features that should be presented by a pervasive computing middleware and how they can aid the development of pervasive applications. Additionally, some pervasive computing

Figure 2. Overview of the chapter
Pervasive Computing

middleware solutions are presented. We conclude the chapter by summarizing the actual state of pervasive computing research, and also discuss possible future directions.

A FIRST GLANCE AT PERVERSIVE COMPUTING

Imagine yourself porting a mobile device, like a handheld, while walking through a shopping mall. Now imagine you are very interested in having a cappuccino. You think it is a pity there is no place in this mall offering cappuccinos. Fortunately, your handheld “knows” that you like cappuccino, and it becomes aware that the shopping mall has just opened a coffee shop. And guess what? Cappuccinos are sold there. Based on this information, your mobile device notifies you about this news, and now you can have your desired cappuccino. While you are savoring it, you are reminded of that book you are interested in. Without hesitation, you take your device out and check to see if any store in the mall has such a book to offer. When the search is finished, you find out that two bookstores are selling the book you want. The search returns all the information you need, such as the price of the book in both stores, discounts, and payment options. With such information at hand, you select the offer best suited for you and request the book. From your mobile device, you provide all the information to complete the purchase. Now, all you need to do is go to the bookstore and get your brand new book.

Wonderful, is it not? Just by porting a mobile device you were able to savor a cappuccino and buy the book you wanted. And both tasks were completed in a very natural way, as if computing had been fully woven into our lives. This is a typical example of a pervasive computing scenario. First introduced by Mark Weiser in his seminal paper (Weiser, 1991), pervasive computing is part of an evolution chain consisting of distributed systems and mobile computing (Satyanarayanan, 2001). It envisions a world where computing and applications are embedded in everyday objects. Clothes, televisions, air conditionings, and cars are examples of such objects. They will be capable of seamlessly interacting with each other in order to perform tasks on behalf of the users by taking intelligent actions or making available relevant information at the right place and at the right time.

Weiser affirmed that pervasive computing can be achieved through three major technologies: cheap and low-power devices, a network infrastructure for communicating these devices, and pervasive applications. At the time this was said, hardware technology was not fully available to support pervasive computing. Wireless networking, as we have today, was neither available nor deployed in mobile devices. Consequently, pervasive applications could not be developed.

This started to change with the introduction of more powerful mobile devices, such as the current smart cellular phones and handhelds, that allowed for the development of more complex applications for such devices. Also, the embedding of wireless networking technologies, like Bluetooth (Bray & Sturman, 2000) and Wi-Fi (Reid & Seide, 2002), on mobile devices has promoted the availability of mobile applications. These technologies have permitted us to give the first steps toward the vision of pervasive computing. This has caused a rush for the first solutions in the field, and many works with this purpose have been developed. Oxygen (http://www.oxygen.lcs.mit.edu), Aura (Garlan, Siewiorek, Smailagic, & Steenkiste, 2002), Smart Space (Stanford, Garofolo, Galibert, Michel, & Laprun, 2003), Portolano (Esler, Hightower, Anderson, & Borriello, 1999), and Wings (Loureiro, Bublitz, Barbosa, Perkusich, Almeida, & Ferreira, 2006) are some examples of works related to the branch of pervasive computing.

However, a long road is still ahead. Despite the hardware advances in the last years, there are still a new set of problems associated with software systems for pervasive environments. For pervasive
computing to become a true reality, applications need to have full access to the information about the users and the environments in which they are situated. This is in a broad sense what has been named context, although many variations for the concept exist nowadays. The current lighting condition, temperature level, and the number of users around a mobile device are some examples of the information associated with the word context. The great challenge that remains within this scope is how to model context information, and mainly how to effectively exploit it. The effective use of context information is one of the key issues to achieve Weiser’s vision of invisible computing (Satyanarayanan, 2001). Still, acting on behalf of users requires pervasive systems to be ready for changes in their interests. Changes in the local and remote resources available should also be considered, as they are important for achieving such pervasiveness.

Going down to the networking level, we find that mobility and host discovery are two important features for pervasive environments. Whereas the former allows embedded applications to perform their tasks uninterruptedly; that is, even when the user is moving through different networks, host discovery permits a device to discover network hosts, and also to be discovered by them. Due to the discovery of such hosts, a device is then able to query for the information and resources they share, informing the user about the most relevant ones.

It is clear the preceding challenges need to be first well understood and solved. Only then can dependable pervasive systems emerge. Therefore, from this point on, we start delineating such challenges, as well as mentioning some of the current solutions for them. However, if any of the introductory ideas are not clear enough, the reader can refer to the literature presented in Table 1.

Table 1. The introductory literature that has been used and their main contributions

<table>
<thead>
<tr>
<th>Reference</th>
<th>Main Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Weiser, 1991)</td>
<td>Overview of pervasive computing</td>
</tr>
<tr>
<td>(Weiser, 1993)</td>
<td></td>
</tr>
<tr>
<td>(Saha &amp; Mukherjee, 2003)</td>
<td></td>
</tr>
<tr>
<td>(Satyanarayanan, 2001)</td>
<td>Challenges brought on by pervasive computing</td>
</tr>
<tr>
<td>(Weiser &amp; Brown, 1995)</td>
<td>The concept of calm technology</td>
</tr>
<tr>
<td>(Garlan, Siewiorek, Smailagic, &amp; Steenkiste, 2002)</td>
<td>Some current solutions for pervasive computing</td>
</tr>
<tr>
<td>(Stanford, Garofolo, Galibert, Michel, &amp; Laprun, 2003)</td>
<td></td>
</tr>
<tr>
<td>(Esler, Highower, Anderson, &amp; Borriello, 1999)</td>
<td></td>
</tr>
<tr>
<td>(Loureiro, Bublitz, Barbosa, Perkusic, Almeida, &amp; Ferreira, 2006)</td>
<td></td>
</tr>
</tbody>
</table>
where we indicate the contribution of references used throughout the first two sections.

PERVASIVE NETWORKING

Pervasive networking is about the plumbing involved in the communication of devices in pervasive computing environments. Therefore, studies within this area range from the design and energy consumption techniques of wireless interfaces to the development of high level protocols, such as routing and transport ones. At this high level, mobility and host discovery play fundamental roles as enablers of pervasive environments. Research in these areas has considerably advanced, and as a result, some practical solutions are already available today. Therefore, in this section we present a review of the concepts associated with mobility and host discovery.

Mobility

You probably receive your mail at your residence, right? Now, consider that you are moving to a new house. Among other concerns, you would probably want to change the mailing address associated with correspondences like your credit card bill. In this case, you must notify your credit card company that you have just moved, and that consequently your mailing address has changed. Either you do this or your credit card bill will be delivered to the old address, which is not a desirable situation.

A scenario similar to the above one is basically what happens in computing environments enhanced with mobility. In other words, mobility must allow a device to change its physical location and still be capable of receiving network packages from the other hosts. Note that, by physical location, we are referring to the network a device is connected to. Therefore, moving through different networks is what requires a node to have its address changed.

Mobility is certainly a fundamental element for pervasive environments. The possibility for providing users with on the move networking enables applications to work in the background by invisibly searching for some relevant content. However, the use of mobility in computing systems inherently leads them to face a set of new and challenging problems, which can be grouped in the following way (Satyanarayanan, 1996):

- **Resource poverty of mobile computing devices:** It is a fact that mobile devices are resource-poor when compared to personal computers. Processor speed and memory/disk capacities are considerably higher in static computers than in mobile ones. Therefore, software for mobile computing need to be well designed in order to save processor usage and storage space.

- **Energy restrictions:** Static computers are plugged to some energy network, which is theoretically an unlimited source of energy. Mobile devices, on the other hand, depend on limited capacity batteries. Therefore, techniques for saving energy should be applied in mobile applications.

- **Variability of wireless links:** Wireless connectivity is still highly variable in terms of performance and reliability. Whereas some buildings provide high-bandwidth and reliable wireless connections, others may provide considerably less bandwidth and reliability. This can be even worse in an open environment, where connection may be shared by lots of users. Undoubtedly, these changes in wireless connectivity need to be addressed in pervasive computing systems by, for example, implementing some network congestion control algorithm.

- **Security of wireless connections:** Due to the broadcast nature of wireless links, they are easier to eavesdrop with than wired ones. Therefore, if security is already an important
feature of fixed networks, for wireless ones it is an even more important feature.

The Degrees of Mobility

The different degrees of mobility have a direct impact over the topology of a network. The more mobile are the network nodes, the more flexible the network needs to be. In the case of Ethernet networks, for example, nodes are too static. Therefore, only in sporadic situations is it necessary to change the network address of a node. Consequently, the network topology does not necessarily need to be flexible. In this case, protocols like DHCP (dynamic host configuration protocol) seamlessly solve the problem of delivering new network addresses to nodes. At the other extreme, a network may be populated by highly mobile nodes. Such a level of mobility allows users to move around areas that, for various reasons, have no fixed network coverage. In these situations infrastructureless networks are more appropriate. That is, nodes should be capable of establishing connections with each other whenever needed. In this case, the network would be formed opportunistically as more and more mobile devices get together.

Within this context, as Sun and Savoula (2002) have already pointed out, three modes of communication can be distinguished when it comes to the degree of mobility: nomadic, cellular, and pervasive communication. In the first case, no connection is necessary when the device is migrating from one network to another. A typical example of the nomadic communication is a user who uses a notebook for connecting to a network both at work and at home. Note that there is no need to keep network connections while the users are moving from work to their house. Only when getting home should the notebook receive a new address for accessing the network. In the cellular communication mode, the network is organized in cells, where each cell is located adjacent to a set of others. All cells have a central element, which provides connectivity for the nodes within them. Therefore, a mobile device can move through different cells and maintain a connection with their central element, becoming thus accessible even when moving. Current mobile telephony networks are an example of this kind of communication, where the base stations act as the central elements. Finally, pervasive communication can be mainly characterized by the lack of a fixed network infrastructure different from the two previous ones. Therefore, nodes should establish connections directly with each other whenever they come close enough. These features are what characterize the so-called ad hoc networks (Chlamtac, Conti, & Liu, 2003), and will be of great importance in the deployment of pervasive computing environments.

Among the current solutions for mobility, we could cite Mobile IP (Perkins, 1997), GPRS (General Packet Radio System), and Bluetooth. Basically, Mobile IP and GPRS are mobility solutions respectively for IP and mobile telephony networks. Bluetooth, on the other hand, is a standard for short-range and low-cost wireless communication in an ad hoc way. Further description concerning these technologies can be found on the Web sites listed in the Useful URLs section.

Host Discovery

Putting it simply, host discovery is about finding other hosts in the network, and also being found by them. This apparently simple concept is of great importance for pervasive computing environments, and can be found in technologies such as Bluetooth and UPnP (Universal Plug and Play). As an example of its usage, consider the acquisition of context information in decentralized environments, such as the available services. By using host discovery, a device can, for example, find the available hosts in the environment and query them for the services they provide. This
Host discovery can be performed either by using a notification-based approach or a query-based one. In the former, a host is discovered when it notifies itself to the others. This requires the host to send its advertisement through the network, so that the others can be aware of it. Such advertisements contain information such as the address, name, and description of the hosts. The advertising task can be executed a single time (e.g., when the host joins the network) or periodically (e.g., each second). The notification-based host discovery is illustrated in Figure 3(a). On the other hand, the query-based approach, illustrated in Figure 3(b), is based on sending discovery messages and waiting for their responses, which contain information about the discovered hosts. Therefore, by retrieving the information contained in these responses, a host is able to contact the discovered hosts.

**CONTEXT IN PERSVASIVE COMPUTING**

A fundamental functionality of pervasive computing applications is to present users with relevant information or services at the right place and in the right time, in a seamless way. Such information can be, for instance, a landmark for tourists to visit based on their preferences. In this process, two key inputs are involved: the needs and interests of the user and the information available both in the environment and in their devices. The former allows the applications to define what sort of information would be relevant to the user. The latter is the source from where such information will be retrieved. Let us get back to our first example, the one presented in the First Glance on Pervasive Computing section. In that case, your desire for a cappuccino and the book you wanted to buy were your needs and interests. Whereas the former could be acquired, for instance, by keep-
ing historic information of your preferences, the information about the book has been explicitly provided by you. Based on both information, the application running in the handheld was able to determine what would be considered relevant information for you. By gathering information from the environment, the application decided that the opening of a new coffee shop and the bookstores in the surrounding area were relevant enough information to present you.

**A Definition of Context**

The preceding discussion should provide at least a first impression of what context really means. In pervasive computing literature, context has been defined in a number of ways. Some researchers have defined context by categorizing the different information associated with it. Gwizdka (2000), for example, identifies two types of context: *internal* and *external*. Internal context provides information about the state of the users, such as their current emotional state. External context, on the other hand, describes the environment on which a user is immersed, for example, informing about the current noise or temperature level. In the work of Petrelli, Not, Strapparava, Stock, and Zancanaro (2000), two types of context are identified: *material* and *social*. Material context is associated with location (e.g., at home), devices (e.g., a handheld, a cellular phone) or the available infrastructure (e.g., available networks). Social context, on the other hand, encapsulates the information about the current social state of the user, for example, in a meeting or a movie theater. Another work, by Schilit and Theimer (1994), defines three categories for grouping context information: *computing context*, *user context*, and *physical context*. A refinement of these categories is presented by Chen and Kotz (2000), through the addition of a fourth category, *time context*. The information associated with each category is presented as follows.

- **Computing context**: Network bandwidth, the cost involved in communication, and available resources, such as printers, displays, and workstations.
- **User context**: People in the vicinity and the location, profile, and current social situation of the user.
- **Physical context**: Lighting and noise levels, current temperature, and traffic conditions.
- **Time context**: Time of the day, day of the week, month and season of the year.

Note that the above ideas do not really define what is context, but instead try to give it a meaning by enumerating the sort of information that could be related. The problem in defining context in this way is that it may be hard to affirm whether some information can be considered context information or not. Additionally, a more general definition of context would certainly enable a better understanding of its role in pervasive computing. Therefore, for the purposes of this chapter, we consider a context as defined by Dey (2001, p. 45).

Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.

**Context Aware Applications**

Considering the current context for determining the actions that can be taken is very natural for us. We commonly use information such as the place we are at and the people around us to guide our actions. When we are in a movie theater, for example, we know how bothersome it is to speak loudly, and so most of us generally do not do so. In this scope, applications that make use of this kind of information are called *context aware ap-
applications. Such applications, however, require functionalities for acquiring and interpreting the context information, and to choose and execute an action based on it. More precisely, three elements are involved throughout this process: context acquisition, context representation, and context reasoning.

Context acquisition concerns the way context information is obtained, namely sensed, derived, or explicitly provided (Mostéfaoui, Rocha, & Brézillon, 2004). Sensed context information is gathered from physical sensors, such as lighting or temperature ones. Derived context information is computed on demand, like the time of the day and the number of people around. Finally, when context information is explicitly provided, the user is responsible for providing it. An example of this acquisition can be viewed on applications which provide a form for users to fill in with their preferences (i.e., favorite kinds of books, movies, food, and entertainment).

Once acquired, context information needs to be made available to the interested applications. This implies that it must be represented in an agreed format, so that the interested applications can “understand” the information they received from the providers. As already pointed out by Held, Buchholz, and Schill (2002), the representation of the context should be structured, interchangeable, composable/decomposable, uniform, extensible, and standardized. Structuring is important for enabling applications to filter pieces of information from the context. Interchangeability is related to the possibility of applications to exchange context information with each other. Therefore, in order to provide this characteristic, a context representation must be serializable. Composition and decomposition enables to compose and decompose context information from different sources. This allows transferring only sub-parts of the information, for example, when it has been updated, to avoid sending the whole context representation. Uniformity claims that different kinds of context information (e.g., user’s profile, device profiles, resource profiles) should be represented in a similar manner in order to ease interpretation by the applications which use them. As the number of terms and variables of a context is difficult to predict, even in quite restricted domains, extensibility is also a fundamental characteristic for context information representation. Finally, as devices and applications can come from different vendors, context information must be based on standards. This would certainly improve the exchanging of context information among pervasive computing applications. Current solutions for pervasive computing represent context in different ways; for example, using key-value pairs, XML documents (Boyera & Lewis, 2005; Ryan, 1999), object-oriented models (Henricksen, Indulska, & Rakotonirainy, 2002), and ontology-based models (Chen, Finin, & Joshi, 2003; Henricksen, Livingstone, & Indulska, 2004; Masuoka, Labrou, Parsie, & Sirin, 2003).

Considering that the context information is represented in a way that applications understand, it is possible to make use of this information and perform context reasoning. Basically, context reasoning is the use of contextual information for guiding the actions an application will take. As Satyanarayanan (2001) has already pointed out, the effective use of context information is a fundamental element in pervasive computing, as a means for achieving the invisibility feature envisioned by Weiser. The context reasoning mechanism of a pervasive computing system can be as simple as if-then-else statements, or as complex as rule-based (Nishigaki, Yasumoto, Shibata, Ito, & Higashino, 2005) and case-based methods (Ma, Kim, Ma, Tang, & Zhou, 2005). An important characteristic of context reasoning systems is the ability to deal with uncertain context information. As the information acquired by sensors (i.e., sensed context information) is prone to errors, applications should consider the quality of the acquired context information when performing their reasoning tasks. To this end, different approaches have been proposed using,
for example, Bayesian networks (Gu, Pung, & Zhang, 2004a) and Fuzzy logic (Ranganathan, Muhtadi, & Campbell, 2004).

DEVELOPMENT OF PERVERSIVE COMPUTING SYSTEMS

Based on our discussion until now, it is easy to realize that the intrinsic features of pervasive computing have an impact on the way software is designed and developed. For example, adaptability, customization, and context sensitivity are some of the characteristics that are constantly associated with pervasive computing systems (Raatikainen, Christensen, & Nakajima, 2002). Different software engineering techniques have been used when dealing with them. In this way, we will now review some of these techniques, as well as how they can be applied in pervasive computing systems. More precisely, we will discuss how the component and plugin-based approaches can be used to provide such systems with adaptability and customization. In addition, we show how they can be aware of changes in the context, through the generation and notification of events.

The Component-Based Approach

Component-based software engineering addresses the development of systems as an assembly of components. More precisely, its focus is on the development of components as reusable entities, as well as on the maintenance and upgrade of systems through the customization and replacement of such components. The main advantages of this reuse and assembly-based paradigm is a more effective management of complexity, reduced time to market, increased productivity, and improved quality (Crnkovic, 2001).

In a general way, a component is a software implementation which can be executed in a logic or physical device and can be reused in several applications of the same domain (Bachman, Bass, Buhman, Dorda, Long, Robert, Seacord, & Wallnau, 2000). The well-defined interface of the component describes the services or events that implement its functionalities. Such interface enables encapsulation of the component’s functionalities, reducing the coupling among them, and also improving the flexibility of the software design.

Figure 4. Component-based architecture
Because the components must be connected to assemble the application, it is necessary software to ensure the interaction among components, managing their service and event dependencies. Generally, such entity is implemented as a software framework. To guarantee the components will behave as expected by the framework, some interfaces, called contracts, are defined (see Figure 4). These contracts, which components are forced to implement, guarantee that the development of independent components satisfies certain standards, allowing the framework to use such components without being aware of their internal implementation details (Bachman et al., 2000).

The Component-Based Approach in Pervasive Computing

Within the scope of pervasive computing, the application of the component-based approach is straightforward in the development and maintenance of software. Due to the dynamics and heterogeneity they present (e.g., different services available, several kinds of protocols, devices with different processing power, and storage capabilities), the reuse and flexibility characteristics of components are mandatory in pervasive computing software design. The combination of these features provides an efficient way for enabling an application to seamlessly adapt. This can be performed either by dynamically changing a component by an equivalent, or by assembling a new functionality to the application.

In this context, reuse is important, for example, due to the increasing number of technologies related to pervasive networking, such as Bluetooth, UPnP, Zeroconf (Guttmann, 2001), and JXTA (Gong, 2001). Since such technologies are based on standard protocols, they can be implemented as software components in order to be reused in several applications. Reuse is not only interesting for communication technologies. Components can also be implemented for many other purposes, such as audio and video streaming and context information retrieval, and yet provide the same reuse feature.

Examples of pervasive computing systems which make use of the component approach include Aura (Garlan, et al., 2002), Runes (Costa, Coulson, Mascolo, Picco, & Zachariadis, 2005), and PCom (Becker, Handte, Schiele, & Rothermel, 2004).

Plug-in-Based Architectures

Applications based on the plug-in approach are characterized by having a functional core with well-defined hooks where extensions (i.e., plug-ins) can be dynamically plugged (see Figure 5) (Mayer, Melzer, & Schweiggert, 2002). The functional core contains only the minimum set of functionalities the application needs to run. Plug-ins, on the other hand, are intended to enhance the application by adding features to it. Therefore, plug-in-based applications can be executed even when no extensions have been installed. Besides, features that are not in use can be safely removed, by plugging out the associated plug-in.

A more revolutionary view of plug-in-based architectures is to consider everything as a plug-in. In this new form of plug-in architectures, the application becomes a runtime engine for managing each plug-in’s life cycle. As a consequence, end user functionalities are entirely provided by means of plug-ins. For such kinds of application, the extension of plug-ins through other plug-ins is thus a fundamental feature (Birsan, 2005).

Plug-in-Based Architectures in Pervasive Computing

The application of the plug-in approach in pervasive computing systems provides them with the needed characteristic of customization. From minimum, but functional software, users can gradually download specific plug-ins to their daily activities, choosing the ones which best supply their needs. Take as an example an envi-
Pervasive Computing

Figure 5. General view of a plug-in-based architecture

Pervasive Computing

It is interesting to note how the pure plug-in approach fits well when applied in pervasive computing. The plug-in runtime environment, obviously equipped with other features, like context sensitivity, can be mapped to a pervasive computing infrastructure. Applications would then be viewed as plug-ins, which could be extended by other plug-ins, and so on. Therefore, plug-ins, in this case, applications, could be installed in the user’s device on demand, and be removed from it when no longer needed.

Within this scope, the fact that plug-ins can be removed without affecting its host application is also important for pervasive computing. Mobile devices are restricted concerning disk and memory capacity, and thus it would be helpful to remove non-used plug-ins in order to save some space.

Practical examples concerning the usage of the plug-in concepts in pervasive computing can be found in middlewares like Wings (Loureiro, et al., 2006), BASE (Becker, Schiele, Gubbels, & Rothermel, 2003), ReMMoC (Coulson, Grace, Blair, Duce, Cooper, & Sagar, 2005), and Plugin-ORB (dAcierno, Pietro, Coronato, & Gugliara, 2005).

Event-Based Systems

An event-based system is the one in which the communication among some of its components is performed by generating and receiving events. In this process, initially a component fires an event,
and after that, such an event will be delivered to all the components interested in it. In an event-based system, a component can assume the role of producer, consumer, or both. The producer is responsible for generating and firing events. The consumer, on the other hand, is a component which registers itself for the occurrence of a particular event, and is notified when such event occurs.

The process of event notification can be performed in two ways: through the event-based or the callback cooperation models (Fiege, Mühl, & Gärtner, 2002). In the former, a key element is the event service. Such an element is responsible for receiving an event from a producer and forwarding it to consumers. To exemplify this process, we have presented an example, which is illustrated in Figure 6. In such a figure, we have six components (A, B, C, D, E, and F), the event service, and two kinds of events (X and Z). Some components act only as producers (components A and C) whereas others only as consumers (components B, D, and F). Finally, component E acts as both a producer and a consumer. The arrows in the figure indicate the flow of event announcements and notifications within the system. In the callback model, on the other hand, the consumer subscribes directly to the provider of the event. Therefore, the provider must keep track of the consumers for each event it provides. When some component is interested in various events, it must thus subscribe to each component providing the event. The so-called Observer design pattern (Gamma, Helm, Johnson, & Vlissides, 1995) is an abstraction of the callback event cooperation model.

The Event-Based Approach in Pervasive Computing

In the context of pervasive computing, the event-oriented approach can be very helpful in the no-
tification of changes in the context. An example would be a user, located in a shopping mall, who is interested in book discounts. Applications dispersed in the environment could frequently deliver events associated with the discounts available in the mall. Through a handheld, the users could then register their interests on events associated, for example, with books discounts. As soon as one of such events is generated by some application, a centralized event service would be able to deliver the event to the user’s handheld, which in turn would notify the user. The event could then provide the user with information such as the store offering the discounts and the associated books.

Another use of events in pervasive computing is for notifying applications about the battery level of a device. In this scenario, through the callback cooperation model, an application could specify a value associated with the remaining battery (e.g., 5%) and register itself to an event which is fired every time the battery level reaches such a value. In this way, an application could be notified when the device is about to run out of energy, in order to save some critical data before the battery be completely consumed.

The event approach has been successfully used in the development of pervasive software systems. Maybe the best example we can cite is the UPnP specification. Its event mechanism, which uses the General Event Notification Architecture (GENA) (Cohen, Aggarwal, & Goland, 2000), permits the notification of changes in a device, such as the services that have been inserted or removed from it. Other examples include the Scooby (Robinson & Wakeman, 2003) and Solar (Chen & Kotz, 2002) pervasive computing middlewares.

MIDDLEWARE FOR PERVERSIVE COMPUTING

The task of building pervasive computing applications can be too tedious if performed from scratch. In other words, the developer will need to deal with low level networking protocols, the way the context is acquired and monitored, notification of changes in the context, and methods for enabling flexibility in their applications. This, of course, deviates the attention of the developers to tasks that are not the purpose of the application. Instead, they should only concentrate on the application logic, that is, the tasks the application must perform. This is where middleware for pervasive computing comes in.

By providing a high level abstraction for application developers, middlewares can considerably speed up development time and also decrease the number of errors in software implementation (Mascolo, Capra, & Emmerich, 2002). Besides abstracting low level details, an ideal middleware should also provide the developer with robustness, efficiency, and security (Aldestein, Gupta, Richard, & Schwiebert, 2005).

These are, however, characteristics that should be presented by any kind of middleware, targeted at pervasive computing or not. From the discussion we have presented until now, it is clear that pervasive computing middleware should not stop at this point. Adaptability, for example, is part of any application in pervasive environments, and consequently, pervasive computing middlewares should present it. Take as an example the RPC (remote procedure call) protocols they use for invoking remote services available in the environment. Considering the diversity of such protocols (e.g., XML-RPC, RMI, SOAP), the middleware could be faced with situations in which it does not implement the RPC protocol used by a certain service. Middlewares for pervasive computing should be capable of transparently overcoming this problem, by downloading the specific RPC protocol implementation, installing it, and finally invoking the service. Moreover, middlewares for pervasive computing should naturally support the development of flexible applications on top of it. In other words, they should provide all the necessary tools for developers to build inherently adaptable
applications. Based on our previous discussions, the component and plug-in-based approaches can be very effective, both for the middleware and the applications running over it.

Another characteristic is that pervasive computing middlewares need to provide a way for applications to retrieve information from the context, and also to be notified about changes related to it. In middlewares for traditional distributed systems this was not required, as the context on which they were executed was too static. Therefore, the sporadic changes in the context could be easily managed by the middleware, with no need to be exposed to the applications. As a result of the dynamics involved in pervasive environments, middlewares can not efficiently take decisions on behalf of applications (Mascolo et al., 2002). It is more reasonable, thus, to expose the context to applications and let them take their own actions. At this point, context representation and event notification are very useful for pervasive computing middlewares.

**Pervasive Computing Middleware Solutions**

Current solutions for pervasive computing middlewares have been focused on different aspects, such as service discovery, service composition, context sensitivity, and networking heterogeneity. These solutions range from general purposes middlewares, like Jini (Waldo, 1999) and Aura (Garlan, et al., 2002), to application domain specific ones, such as healthcare (Bardram & Christensen, 2001) and e-learning (Apostolopoulos & Kefala, 2003). Due to their importance in the development of pervasive computing applications, in this section we outline some of these solutions.

**Jini**

Jini is a middleware focused on service discovery and advertisement. Jini services are advertised in service catalogs through a Java interface. When the service catalog address is known a priori, Jini clients make requests directly to the service catalog in order to discover services. The service catalog returns a proxy for each discovered service, which is used by clients for remotely invoking it. If the service catalog address is not known, requests are performed in order to find it. As the service provider—and consequently its services—may not be available in the network, Jini implements a lease mechanism to help clients avoid finding unavailable services. Therefore, when providing a service, its provider receives a lease, which must be renewed at a specified time interval. If the lease is not renewed, the service is removed from the service catalog, and thus can no longer be found by Jini clients.

**OSGi**

The open services gateway interface (OSGi) is a specification supporting the development of Java service-based applications through the deployment of components known as bundles (Lee, Nordstedt, & Helal, 2003). A major advantage of the OSGi specification is that such bundles can be installed, uninstalled, and updated without the need to stop and restart the Java applications. In the scope of pervasive computing, this is a fundamental feature, as it enables pervasive computing systems to adapt themselves in a completely transparent way to their users.

The main idea behind OSGi is the sharing and discovery of services. In other words, bundles are able to advertise and discover services. When advertising a service, a bundle can define a set of key-value pairs, representing the service’s properties. Such properties can thus be useful for other bundles, in order to discover the services they need. These advertisement and discovery processes are both performed through a registry, managed by the OSGi implementation. In this way, it is able to keep track of all services cur-
rently advertised and being used, thus enabling a bundle to be updated even when other bundles are executing and using its services.

**RCSM**

Reconfigurable context-sensitive middleware (RCSM) (Yau, Karim, Yu, Bin, & Gupta, 2002) addresses context-awareness issues through an object-based framework. The context-independent information of an application is implemented in programming languages such as C++ and Java. The context-sensitive information is implemented as an interface, using the context-aware interface definition language (CA-IDL). This interface has a mapping of what actions should be executed according to each activated context. In this way, the application logic is isolated from the context specification.

**SOCAM**

Service-oriented context-aware middleware (SOCAM) (Gu, Pung, & Zhang, 2004b) supports the development of context-aware services where ontologies are used for representing context information. There are two types of ontologies: *high-level ontologies* describe generic concepts which are domain-independent, such as person, activity, location, and device, and *domain-specific ontologies* define concepts which concern specific domains, such as vehicle and home domains.

**Aura**

The main focus of Aura is to minimize the distraction of users by providing an environment in which adaptation is guided by the user’s context and needs. This project, developed at the Carnegie Mellon University, has been applied in the implementation of various applications, such as a wireless bandwidth advisor, a WaveLan-based people locator, an application which captures the user’s intent in order to provide task mobility among different environments. All these applications use components provided specifically by the Aura project, as well as other components, such as the service registry functionalities provided by Jini and Coda (Satyanarayanan, 2002).

**CURRENT AND FUTURE TRENDS IN PERSERIVE COMPUTING RESEARCH**

Undoubtedly, research in the pervasive computing field has considerably advanced. At the networking level, for example, wireless communication is already possible through technologies like Bluetooth, Wi-Fi, and Zigbee. Wi-Fi, although providing a data rate of up to 54 Mbps, has high energy consumption. Bluetooth and Zigbee consume considerably less energy than Wi-Fi, but provide less transmission rates, respectively of 1 Mbps and 250 Kbps at most. The tradeoff of these characteristics must then be analyzed when deploying a pervasive computing environment. Despite these advances, enhancements must still be achieved in this area. Among them, we could point out the power consumption of wireless interfaces. Are current solutions well suited for today’s appliances? Can power consumption be improved? Or better, can the relation between power consumption and transfer rate of a wireless interface be improved? Considering this variability of power consumption and transfer rate in wireless interfaces, what would be an ideal configuration for the network and the appliances (i.e., the set of wireless interfaces in the environment and the devices) to minimize power consumption and maximize the transfer of data?

Mobility and ad hoc networking protocols have also been developed, like Mobile IP, GPRS, UPnP, and Zeroconf. The mix of wireless interfaces and these protocols have already leveraged the deployment of some experimental pervasive
environments. One point to be considered in this scope is the degree of pervasiveness of such solutions. In other words, how seamless are they concerning configuration, initialization, and finalization? Do we need to start and stop them all the time? One could still think about the performance of these solutions in current mobile devices. Do they consume too much energy? If yes, how can they be improved?

When considering the context-awareness, undoubtedly some efforts have been made in specific application domains. As examples, we could cite tourist information, healthcare, sports, learning, multimedia, and intelligent houses. However, context is far from being used as it is intended. For example, some of the current solutions focus on discovering needed services, but do not well address the assembly of new functionalities based on the requirements of the users and the resources available. In this case, what sort of methods should be applied in order to improve this assembly? Is the context information represented in a reasonable way, so that it simplifies such assembly? Work in this area should involve more precise representation of the context, including the resources and the interests of the user, as well as more intelligent methods for better determining what is relevant to the users, and what is not. Determining what is relevant to the user also requires the application to be capable of capturing the user’s preferences, needs, and different states (i.e., social, emotional, physical). User preferences are mostly acquired by providing forms which users fill in. However, how boring can it be for a user to update such forms? Considering that this is a boring task, are they willing to pay this price for some degree of pervasiveness? Concerning the needs of the user, they are commonly obtained explicitly; that is, the users provide to the system what functionality they are interested in. Can this acquisition be enhanced? Is it possible to draw a profile of the user from its reoccurring needs in order to enable pro-activity? What about the information concerning the states of the user? Is it reasonable to require the users to explicitly set them in their personal devices? Can this information be acquired in a different way, through sensors placed in the user’s body for example?

Another point to be analyzed concerns the social impact pervasive computing may cause. Consider, for example, the way people will react to sensors that detect their presence and suggest actions for them. Will they be pleased for these suggestions or angry because there is always some boring gadget telling them what to do? One should consider also whether people will accept having sensors in their bodies for obtaining information like their health and emotional state. Will you? What about privacy and security? Many people are still mistrustful of checking their banking accounts through the Internet, what would they say about sharing personal information in an open environment inhabited by all kinds of people? Certainly, all these social factors are as important as technological ones, and thus, must be deeply investigated.

PERVASIVE COMPUTING FOR KNOWLEDGE AND LEARNING MANAGEMENT

As we have seen throughout this chapter, the evolution that has taken place in hardware and networking technologies, along with the dissemination of mobile devices like cellular phones, make pervasive computing the prevailing paradigm for the next computing systems. Within the scope of this trend, one of the aspects that pervasive computing can greatly contribute is knowledge management, mainly when we consider the increased demand for learning independently of time and place (Lytras, Pouloudi, & Poulymenakou, 2002). The use of information technology (IT) in the context of knowledge management is already commonplace. As already stated by Marwick (2001), many different IT solutions can be used...
Pervasive Computing

for this purpose, with the goal of enabling knowledge to be acquired, created, and disseminated by transforming it from tacit to explicit forms and vice-versa. One of the great advantages of IT, in this case, is the possibility to transpose the barriers of time and space. Based on this, it is not hard to realize how pervasive computing can aid in further transposing such barriers. With pervasive computing technology, members of an organization are able to establish meetings, anytime and anywhere, with the goal of sharing their knowledge. In this scenario, video and audio over the Internet could be used to give the impression that the members are in a real meeting. Therefore, people from organizations geographically distributed would not need to travel to meet each other. Another use of pervasive computing technology is for the generation of reports with the intent of disseminating someone’s knowledge. With the possibility of having Internet connection whenever needed, members of an organization could prepare reports about a specific topic and then distribute them to the others, no matter where they are. Within this process, any needed document, either internal or external to the organization, could also be accessed, enabling knowledge to be acquired, created, and disseminated in a ubiquitous way.

Due to this possibility of extending the limits of knowledge acquisition, creation, and dissemination, it is not surprising to see the first solutions trying to combine pervasive computing and knowledge management. An example is the work of Will, Lech, and Klein (2004), which proposes a tamagotchi-based solution for supporting mobile workers in finding relevant information for the work they perform. Basically, this solution works by enabling a mobile device to interact with information suppliers in a seamless way, through a continuous and proactive matching between the information they provide and the one needed by the mobile workers. Another work in this category is the p-learning Grid, also known as the pervasive learning Grid (Liao, Yang, & Hsu, 2005). As its name indicates, the p-Learning Grid is targeted to support mobile learners in pervasive environments. The overall idea of this work is to represent learning objects (LOs) using Grid services (Foster, Kesselman, Nick, & Tuecke, 2002), distributing them among several connected computers (i.e., Grid infrastructure). In this way, through such an infrastructure, LOs can be dynamically discovered by client devices, using, for example, the current learning needs of users. Other solutions in this field include the work of Amman, Bright, Quirchmayr, and Thomas (2003) and the GetSmart system (Marshall, Zhang, Chen, Lally, Shen, Fox, & Cassel, 2003).

CONCLUSION

In this chapter, we have discussed some concepts surrounding pervasive computing. We have provided the reader with a high level understanding of the pervasive computing paradigm at different levels, and thus we have ranged from networking to software engineering issues. By using this approach, we presented a broad vision related to pervasive computing to aid researchers, students, and professors in research and teaching activities.

Based on the concepts we have discussed, it is possible to conclude that the application of pervasive computing in the real world is still in its beginning. Many efforts have still to be made in order to bring the primary vision of pervasiveness to real life. One could ask whether such vision will really be conceived. This is, for now, a question which is still unanswered. Answering it will require deep theoretical and, mainly, practical studies. This is what researchers should focus on, and thus, this work has been an introductory theoretical contribution to this end. We believe the concepts presented here will be summed up in other works, and also be useful when developing
real world applications, with the intent of reaching a reasonable vision of pervasive computing.

REFERENCES


**ENDNOTE**

APPENDIX I: INTERNET SECTION: UBIQUITOUS COMPUTING GRAND CHALLENGE

The Ubiquitous Computing Grand Challenge is a community of researchers from different parts of the world targeting ubiquitous computing (http://www-dse.doc.ic.ac.uk/Projects/UbiNet/GC). Their research is focused at different levels of ubiquitous computing, from social to technological. With this purpose, this community has been proposing a set of, as they call it, foothill projects, within the scope of ubiquitous computing. This community still provides a mailing list used by their members to discuss the projects and directions of the community, among other adjacent topics. The registration to the list is open.

Interaction

Select one of the foothill projects presented on the Ubiquitous Computing Grand Challenge Web site and prepare one of the following items:

1. A research paper with at least 3000 words, without counting images, or
2. A presentation at least 40 minutes long.

Both the paper and the presentation should present in what way the selected project contributes to the ubiquitous computing research, the state of the art, and practical applications. Remember that it could be helpful to access the mailing list, for example, to generate some interesting discussion concerning the selected foothill project.

APPENDIX II: CASE STUDY

A Usual Day with Pervasive Computing

At half past six in the morning, Jessica’s alarm clock wakes her up, as programmed. As soon as she gets off the bed, the curtains of the bedroom are automatically opened and the alarm clock stops yelling. Within one hour, she dresses for work, takes her breakfast, gets her stuff, and leaves.

When arriving at work, a camera positioned at the entrance of the parking lot recognizes Jessica and her car, and thus, the gates are automatically opened. She greets the gateman and enters. At this point, her car automatically selects the best parking options. She chooses one of them, and leaves the parking lot.

Jessica is about to enter the company when a sensor in the entry of the building detects her presence, and by knowing the elevator she usually takes to go up, warns her that it is on maintenance. Based on the number of people waiting at each elevator, it suggests the best option for Jessica.

Finally, Jessica gets to her office, and at this moment environmental conditions, like lighting or air conditioning levels or curtains opening, are automatically adjusted according to her preferences. Furthermore, a coffee machine, which knows that Jessica usually drinks a cup of coffee in the morning, greets her and asks if she would like some. She has a meeting within a few minutes, but she thinks she has enough time to taste her daily coffee. Then, through a voice command, Jessica orders the machine to prepare it. After finishing the coffee, she then leaves the room for another day of work.

Late in the night, Jessica gets ready to return home. When leaving the parking lot, her car informs her that the usual way she takes home is too congested and it automatically provides an alternative route.
On her way home, she receives a phone call on her cellular phone. As she is currently driving, the phone redirects the caller’s talk to an audio output located in the car. Jessica can use an available microphone to talk to the caller by forwarding the audio streaming to the cellular phone. Therefore, the focus on the driving is not compromised.

When she gets home, the mailbox identifies her and notifies her about the three letters left in it. She takes the correspondences and enters her home, when again the environment is adjusted according to her preferences. Furthermore, the answering machine detects her presence and automatically informs her of missed calls. Through a voice command, Jessica starts listening to each message left. She then finds a message by her mother, asking her to return the call as soon as she gets home. Jessica stops listening to the missed calls and orders the answering machine to dial her mother’s number through her cellular phone. After talking to her, Jessica has dinner and finally gets ready to sleep. She lies down on her bed, and automatically the air conditioning is turned on, the curtains are closed, and the alarm clock is set to wake her up in the morning to have one more usual day with pervasive computing.

Questions

1. What other pervasive computing features would be helpful for Jessica’s daily activities? For each feature provide a detailed description of how it would be fit into the above scenario.
2. Do you think the current society is ready for this kind of technology? Explain your answer.
3. Considering the hardware and software technologies we have today, identify/describe issues of creating the pervasive computing scenario described.

APPENDIX III: USEFUL URLs

UPnP Forum Web site
http://www.upnp.org

Wi-Fi Alliance
http://www.wi-fi.org

Bluetooth Special Interest Group
https://www.bluetooth.org

What is General Packet Radio Service?

Zigbee Alliance
http://www.zigbee.org

Home Networking with Zigbee
http://www.embedded.com/showArticle.jhtml?articleID=18902431

Mobile IP for IPv4
http://www.ietf.org/html.charters/mip4-charter.html

Zero Configuration Networking (Zeroconf)
http://www.zeroconf.org

OSGi Alliance
http://www.osgi.org

APPENDIX IV: FURTHER READING


APPENDIX V: POSSIBLE PAPER TITLES/ESSAYS

- Pervasive computing: a utopia of human mind or a promising paradigm for everyday life?
- Social impacts of pervasive computing
- Bringing pervasiveness to the real world: practical applications of pervasive computing
- Dealing with privacy of context information in pervasive computing systems
- The impact of pervasive computing on software development

This work was previously published in Ubiquitous and Pervasive Knowledge and Learning Management: Semantics, Social Networking and New Media to Their Full Potential, edited by M. Lytras and A. Naeve, pp. 1-34, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 1.19
Pervasive and Grid Computing Merging

Lorenzo Oliveira
Federal University of Campina Grande, Brazil

Emerson Loureiro
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

INTRODUCTION

Nowadays, we are experiencing an increasing use of mobile and embedded devices. These devices, aided by the emergence of new wireless technologies and software paradigms, among other technological conquests, are providing means to accomplish the vision of a new era in computer science. In this vision, the way we create and use computational systems changes drastically for a model where computers lose their “computer appearance.” Their sizes were reduced, cables were substituted by wireless connections, and they are becoming part of everyday objects, such as clothes, automobiles, and domestic equipments.

Initially called ubiquitous computing, this paradigm of computation is also known as pervasive computing (Weiser, 1991). It is mainly characterized by the use of portable devices that interact with other portable devices and resources from wired networks to offer personalized services to the users. While leveraging pervasive computing, these portable devices also bring new challenges to the research in this area. The major problems arise from the limitations of the devices.

At the same time that pervasive computing was attaining space within the research community, the field of grid computing (Foster, Kesselman, & Tuecke, 2001) was also gaining visibility and growing in maturity and importance. More than
just a low cost platform for high performance computing, grid computing emerges as a solution for virtualization and sharing of computational resources.

In the context of virtual organizations, both grid and pervasive computing assemble a number of features that are quite desirable for several scenarios within this field, notably the exchanging of information and computational resources among environments and organizations. The features of these technologies are enabling system designers to provide newer and enhanced kinds of services within different contexts, such as industry, marketing, commerce, education, businesses, and convenience. Furthermore, as time goes on, researchers have made attempts of extracting and incorporating the better of the two technologies, thus fostering the evolution of existing solutions and the development of new applications. On the one hand, pervasive computing researchers are essentially interested in using wired grids to hide the limitations of mobile devices. On the other hand, grid computing researchers are broadening the diversity of resources adhered to the grid by incorporating mobile devices.

This chapter presents part of our experiences in the research of both pervasive and grid computing. We start with an overview about grid and pervasive technologies. Following, there are described and discussed approaches for combining pervasive and grid computing. These approaches are presented from both perspectives of grid and pervasive computing research. Finally, in the last section, there are presented our criticisms about the approaches discussed and our hopes about the future steps for this blend of technologies.

GRID AND PERVERSIVE COMPUTING

Grid computing (in short grids) was born as a low cost alternative for high performance computing, rapidly evolving to a solution for virtualization and sharing of resources. By spreading workload across a large number of computers, users take advantage of enormous computational resources that would otherwise be prohibitively expensive to attain with supercomputers. To day, grids try to leverage a world where the access to any computational resources across sites is standardized, ubiquitous and reliable.

Grids are characterized by the use of heterogeneous and non-dedicated resources, scattered under different administrative domains and linked through low-bandwidth channels. There are no guarantees neither whether a certain resource will be available at a certain moment nor, once available, it will keep available during any interval of time. Due to these features, grids are best suited for applications that can adapt themselves to deal with intermittent resources.

More recently, grids are going towards defining a service-oriented architecture, OGSA, where resources are advertised and used as services. OGSA extends traditional service-oriented architectures by the standardization of a number of helper services, concerned with the maintenance of the whole infrastructure. Such services include load balancing, QoS assurance, redundancy of services, and so forth.

Pervasive computing is also a kind of distributed computing. The general vision is that computation is embedded in objects of everyday life, such as clothes, rooms, automobiles and domestic devices. These objects and environments collaborate on behalf of people, balancing proactivity, context information and pre-known information about users’ profiles to provide personalized services to users, on the right place and on the right moment.

Pervasive computing materializes in the form of smart spaces. These environments are characterized by collaboration between embedded devices, wired infrastructure and users. Normally, users interact with smart spaces through portable devices. These devices carry information about
the users’ profiles and may tell the environment how the user wishes to interact with it. Alternatively, the environment itself may know about users’ profiles, and query them by “sensing” users through sensors. Regardless of how the environment gets aware of users’ profiles and how it senses the presence of users, the basic principle of pervasive computing is to disappear from human perception. In other words, users should benefit from decisions made by environments without being prompted every time about what they wish the environment does for them.

THE PERVERSIVE PERSPECTIVE

Pervasive computing is mainly characterized by the use of resource constrained mobile devices. Thus, from the perspective of pervasive computing, the main research focus is on using resources of grids for supplying lacks of mobile devices.

Current approaches for bringing together grids and pervasive computing will be discussed in the following.

Using Grids for Hiding Limitations of Mobile Clients

In the perspective of pervasive computing, using grids for supplying the lack of resources of mobile devices seems to be the major contribution that grids may bring to the pervasive computing research. Grids have plenty supply of resources that are needed by mobile clients. Nevertheless, current attempts to virtualize the capacity of mobile clients (Clarke & Humphrey, 2002; Hingne, Joshi, Finin et al., 2003; Kumar & Song, 2005;) still falls into strong assumptions about the communication between mobile client and grid. Particularly, successful applications of this approach are yet restricted for those cases where the interaction between mobile client and grid requires low data exchange or where some alternative is found for supplying the system with needed data.

For instance, Kumar and Song (2005) proposed a grid-based infrastructure where mobile clients run only helper applications linked to the grid. As proof of concept, a virtual file system was developed linking some folders in the PDA’s file system to a bigger hard disk in the grid. Thus, files and full programs may be moved to the grid, staying in the PDA only those files and programs more frequently used. All content moved to the grid is consumed on demand through streams.

The approach of Kumar and Song (2005) reveals the main weaknesses of integrating pervasive and grid computing. Due to limitations of battery autonomy, portable devices should avoid using intensively their network interfaces. Therefore, a successful integration between grid and pervasive computing should consider applications where there are not large amounts of data being exchanged between portable devices and grid.

The issues of restricted data exchange between portable devices and grid are better targeted by Hingne et al. (2003). In this approach, the authors consider a niche of applications where data needed by portable clients require intensive processing over large amounts of data. However, the authors assume these data can be obtained from different fonts than users’ devices, such as sensor networks and databases, both directly linked to the grid.

In this scenario, users use portable devices to perform queries for data that need to be transferred, and eventually created, on-demand. Data can be directly gathered from data sources, as a databank or some sensor, or may need to be created from the processing of some other data. In the last case, a grid is used to process data and thus reduce the overall client waiting time.

Besides, on the one hand, the approach of Hingne et al. (2003) presents a solution for the issues discussed before, on the other hand it may be posing an economical challenge. Deploying an entire infrastructure, for example a sensor network, to substitute the mobile device role of acting as the system’s entrance of data may be prohibitively expensive. Also, this solution may
restrict user mobility in the sense that users become restricted to interact with the system only where the infrastructure was deployed.

**Using Grids as Test Bed of Pervasive Applications**

Adequate test beds are used for revealing problems very difficult to identify in traditional development environments. The development of any kind of application achieves better results when prototypes are tested and developed in situations and scenarios close to the real ones. There are many examples of well succeeded test beds for classes of applications, for example PlanetLab, which is an overlay network over the Internet intended to be a test bed of distributed applications, specially, the peer-to-peer ones.

Keeping this in mind, Davies, Friday, and Storz (2004) set up their own approach for getting benefit from grids. The authors underline common characteristics of grids and pervasive computing and use this as a starting point for defining a grid-based test bed of pervasive applications.

The ability of testing and developing pervasive applications in an environment that simulates real environments is certainly a great contribution. Problems that are difficultly detected inside traditional development environments are easier identified and fixed when applications are tested in appropriate environments.

Nevertheless, some questions remain open. For instance, grids are traditionally built over wired networks and smart spaces are composed of wireless devices. In order to try to be closer to real smart spaces, such a grid-based test bed must simulate the network interfaces of mobile clients, including bandwidth, power consumption and coverage area. Another point of investigation is the gap of performance between small portable devices and computers of wired grids. Tests performed in hosts rather powerful than the real devices where the applications should run may hide problems that should be identified in this phase of tests. After all, it is still needed to know whether the tested application is the same one to be deployed in mobile devices. This becomes an issue due to differences in architectures, as well as hardware and software-specific optimizations that may contrast the results given in the test and deployment phases.

**Using Grids to Collect and Process Data from Sensor Networks**

Embedding computation in everyday objects means, most of time, deploying sensors and actuators across environments. Sensors are responsible for bringing to the computational world information about the physical world. Conversely, actuators obtain information from the computational world and are responsible for transforming information into actions in the physical world. Sensors may produce large amounts of data that must be processed and interpreted before an action is taken. Processing large amounts of data generally implies the use of large amounts of computational power. Therefore, grids may play an important role in gathering and processing information from sensor networks (Barratt et al., 2003; Gaynor, Moulton, Welsh et al., 2004).

In this context, Gaynor et al. (2004) discuss challenges and research directions for integrating sensor networks to a standard wired grid. Despite all challenges, this approach seems to balance advantages from both technologies. In fact, this alternative requires not much effort to be applied, since it is a direct application of grids. As posed by Gaynor et al. (2004), data standardization is possibly a challenging task in providing such architecture. Once collected from sensors, it is important to share data in a uniform manner. Thus, different applications and devices may consume data without being overburden with logics for transforming data to an understandable format.
THE GRID PERSPECTIVE

Traditionally, grid computing research has been focusing on high performance. Even with the aim change for sharing any kind of computational resources instead of only CPU cycles and storage space, the grid community is still hungry for more computational power. In this context, the main focus of the grid community is to use mobile devices in the grid for solving computational tasks and storing files. In what follows, approaches for getting benefit of pervasive computing inside the grid research are summarized.

Augmenting Grid Resources through Mobile Devices

The general idea is to use small mobile devices as nodes of high performance grids (Litke, Skoutas, & Varvarigou, 2004; McKnight, Howison, & Bradner, 2004; Phan, Hung, & Dulan, 2002). Nodes connect to the grid as soon as they approximate to some access point and are disconnected from it when they leave the access point coverage area. While connected to the grid, mobile devices can receive and submit tasks.

This may sound a little atypical considering all restrictions of small mobile devices. However, research groups around the whole world bet in the whole number of small devices working together instead of the computational power of one lone device. Phan et al. (2002) define an architecture as including mobile devices into a wired grid. Mobile devices enter the grid through proxies. This proxy runs an appropriate grid middleware, such as Globus or Condor, to publish itself as a node. Mobile devices are not visible to the rest of the grid. They only contribute to enlarge the computational power of the proxy, which is responsible for decomposing and spreading received work across the mobile devices linked to it.

Besides all drawbacks discussed before, this approach has yet the serious problem of how a given task can be broken into smaller tasks for sharing with mobile devices. Definitely, this is a non-trivial problem and may overburden grid users, who will need to provide the means for a proxy node to split a task into small ones.

In the grid perspective, augmenting the number of devices supported by the infrastructure is always a good deal. Both mobile devices and grid have an opportunity to obtain benefits from this merging. Moreover, constant advances in mobile technology trend to reduce the drawbacks of using mobile devices for high processing. Nonetheless, the actual technology still poses several challenges for this usage of mobile devices. In addition to issues already discussed, the mobile nature itself of these small devices may cause loss of work. For instance, when a portable device receives some work to perform and, just before concluding it, the device leaves the access point coverage area. There is also an economic question to be solved. Small mobile devices have clearly great potential advantages of being part of a grid in the sense that a grid may supply it with all the resources such devices may need. However, while part of a grid economy, mobile devices outwears themselves rather more than ‘conventional’ grid resources. Therefore, the grid economic model itself should adapt to address this new class of devices, making this integration profitable for both mobile and grid users.

Mobile/Wireless Grids

Mobile grids have as main argument carrying grids where cables do not reach (Chu & Humphrey, 2004; Kurkovsky & Bhagyavati, 2003; Park, Ko & Kim, 2003). The basic idea is to use high performance where the wired infrastructure is not available and clustering mobile devices is the last choice. In this context, software programs are ready to use resources of wired grids as well as switch to an ‘unplugged’ mode, where mobile devices collaborate to achieve better throughput for distributed processing.
As an example of this approach, Kurkovsky and Bhagyavati (2003) combine *mobile devices* to enable higher processing capacities for small devices. To achieve this, the defined architecture uses resources from an access point, for maintaining information about available devices in range and distribution of tasks among nodes. Any *mobile device* may submit tasks to the mobile grid. For this, the device must ask the access point to coordinate the distribution of tasks, re-submissions and collecting of results.

The idea of using on-demand wireless high processing sounds good, but certainly using a set of *mobile devices* in place of a wired grid does not solve the problem of high processing. However, this may be the only alternative on critical situations (Akogrimo, 2006).

Again, overloading these resource constrained devices trend to overdrive them, reducing thus their autonomy and without gaining anything with this. If we consider situations where portable devices accept tasks whose processing time spends more time then its autonomy, all effort for creating a grid in an ad-hoc fashion will be lost.

**Monitoring Grids and Controlling Applications through Portable Devices**

This is the more simplistic use of *mobile devices* in grids. The approach seems to have no great challenges to be transposed. On the other hand, the benefits obtained are rather limited. The general idea is quite simple. In a wired grid, it is enough to provide means for mobile clients to monitor relevant pieces of the grid, such as the grid healthy and application progress (Srinivasan, 2005; Taylor, Shields, Wang & Philp, 2003). Alternatively, users can somewhat control the grid operation, changing the way the work is doing.

The clear advantage of this approach is the low burden over portable devices. This approach requires low interaction and low data exchange between mobile clients and grid. There are also restrictions about the kind of data to be monitored. If large amounts of data are needed to be transferred to *mobile devices* then this approach may be inadequate owning to reasons already discussed in this chapter. This approach may be unsuitable in those scenarios where users want real-time access to monitoring data. Depending on the size of data, displaying a continuous stream of data may overdrive *mobile devices*. (Taylor et al., 2003) also relates to this approach. In this work Taylor et al. (2003) use a grid to simulate the formation of galaxies. *Mobile devices* may be used to monitor the workflow of grid services involved in the process.

**CONCLUSION**

In this chapter we discussed various approaches for combining grid and pervasive computing. The approaches were grouped into categories that were exemplified and criticized. The list is not exhaustive, but we believe it is sufficiently representative of how research groups are attempting to blend the technologies.

Resource constraints of *mobile devices* still are the bottleneck for this integration. In the perspective of pervasive computing, massive search applications (e.g., cryptographic key breakers) are the best example of class of application with all features for a successful integration between grids and pervasive computing—small pieces of data requiring intensive processing that produce more small pieces of data.

All efforts presented here are valid and, in well defined scenarios, they show the technologies can be successfully integrated. The approaches discussed in this chapter have their own application niche and, hence, a space where they can evolve. Nevertheless, our hopes are that while small *mobile devices* were so resource constrained, difficultly researchers will have both freedom and tools enough for developing their ideas. On the other hand, there are several problems
related to both grid and pervasive computing research: large-scale deployment, distributed resource management, economic models, interface standardization, heterogeneity, interoperability, scalability, runtime adaptability, fault tolerance, service composition, service discovery, security, resources semantic, and so on. Hence, while all the issues pointed out in this chapter were yet challenging tasks, both grids and Pervasive computing may contribute to each other by exchanging experiences.

REFERENCES


Barratt, C., Brogni, A., Chalmers, M., Cobern, W.R., Crowe, J., Cruickshank et al. (2003). Extending the grid to support remote medical monitoring, Proceedings of the 2nd UK e-Science All Hands Meeting (AHM’03), Nottingham, United Kingdom.


KEY TERMS

Context Information: Any relevant information regarding the environment and its users. Smart spaces can use context information to deliver personalized services to users.

Embedded Devices: An embedded device is a special-purpose computer system, which is completely encapsulated by the device it controls. Examples of embedded devices include home automation products, like thermostats, sprinklers and security monitoring systems.

Grid Services: A kind of web service. Grid services extend the notion of web services through the adding of concepts regarding the WSRF specification, such as statefull services.

Intermittent Resources: Both in a grid and smart space, intermittent resources are any kind of resource, as computers or services, which may appear or disappear at any time.

Mobile Devices: Any low-sized portable device used to interact with other mobile devices and resources from smart spaces. Examples of mobile devices are cellular phones, smartphones, PDAs, notebooks and tablet PCs.

Peer-to-Peer: Architecture for distributed applications without central control where each participant acts both as client and server.

Sensor Networks: Network of many, spatially distributed devices using sensors to monitor conditions at different locations, such as temperature, sound, vibration, pressure, motion or pollutants.

Service-Oriented Architecture: Collection of services communicating with each other. Services are self-contained and do not depend on the context or state of other services.

Test Bed: In the context of software engineering, an execution environment configured for testing software.

ENDNOTES

1 Open grid service architecture
2 Quality of service

This work was previously published in the Encyclopedia of Networked and Virtual Organizations, edited by G.D. Putnik and M.M. Cunha, pp. 1223-1229, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
ABSTRACT

The advancement of technologies to connect people and objects anywhere has provided many opportunities for enterprises. This chapter will review the different wireless networking technologies and mobile devices that have been developed, and discuss how they can help organizations better bridge the gap between their employees or customers and the information they need. The chapter will also discuss the promising application areas and human-computer interaction modes in the pervasive computing world, and propose a service-oriented architecture to better support such applications and interactions.

INTRODUCTION

With the advancement of computing and communications technologies, people do not have to sit in front of Internet-ready computers to enjoy the benefit of information access and processing. Pervasive computing, or ubiquitous computing, refers to the use of wireless and/or mobile devices to provide users access to information or applications while the users are on the go. These mobile devices can be carried by the users, or embedded in the environment. In either case, these devices are connected, most likely through a wireless network, to the Internet or a local area network (LAN).

Mobile technologies come in a large variety and are ever changing. In order to gain the business value of pervasive computing, and at the same time keep the supporting cost under control, it is important to develop an architecture solution. A service-oriented architecture (SOA) would allow an enterprise to easily provision functions to be accessible by certain types of pervasive channels. A service-oriented architecture would also make it possible to quickly integrate data generated by pervasive devices and make them available in the form of an information service.

In this chapter, we will first look at the communication networks and mobile devices that create the various information-access and information-generation touch points in a pervasive...
computing environment. Then we will discuss the applications and interaction models for pervasive computing. Finally, we will describe a service-oriented architecture that an enterprise can adopt in order to effectively and efficiently support pervasive computing.

MOBILE COMMUNICATION NETWORKS

Mobile communication technologies range from personal area networks (PANs; a range of about 10 meters) and local area networks (a range of about 100 meters) to wide area networks (WANs; a few kilometers). From a network-topology perspective, most networks are based on a client-server model. A few are based on the peer-to-peer model.

Wireless PANs

A wireless personal area network allows the different devices that a person uses around a cubicle, room, or house to be connected wirelessly. Such devices may include the computer, personal digital assistants (PDAs), cell phone, printer, and so forth.

Bluetooth is a global de facto standard for wireless connectivity (Bluetooth SIG, 2005). The technology is named after the 10th-century Danish King Harald, who united Denmark and Norway and traveled extensively.

HomeRF is an early technology for wireless home networking, first marketed in 2000.

The Institute of Electrical Engineers (IEEE) 802.15 wireless-PAN effort (IEEE, 2005a) focuses on the development of common standards for personal area networks or short-distance wireless networks. One technology out of this effort is ZigBee, which is based on the IEEE 802.15.4 standard.

ZigBee is a low-cost, low-power-consumption, wireless communication-standard proposal (ZigBee Alliance, 2005). Formerly known as FireFly, ZigBee is being developed as the streamlined version of HomeRF. A streamlined version would allow most of the functionality with less integration and compatibility issues.

ZigBee’s topology allows as many as 250 nodes per network, making the standard ideal for industrial applications. Radio-frequency-based ZigBee is positioned to eventually replace infrared links. To achieve low power consumption, ZigBee designates one of its devices to take on the coordinator role. The coordinator is charged with waking up other devices on the network that are in a sleep mode, moments before packets are sent to them. ZigBee also allows coordinators to talk to one another wirelessly. This will allow for opportunities for wireless sensors to continuously communicate with other sensors and to a centralized system.

For enterprise computing, the wireless PANs are within the corporate firewall. They do not

<table>
<thead>
<tr>
<th>Technology</th>
<th>Radio Frequency</th>
<th>Maximum Distance</th>
<th>Data Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bluetooth</td>
<td>2.4 GHz</td>
<td>10 meters</td>
<td>721 Kbps</td>
</tr>
<tr>
<td>HomeRF</td>
<td>2.4 GHz</td>
<td>50 meters</td>
<td>0.4-10 Mbps, depending on distance</td>
</tr>
<tr>
<td>ZigBee</td>
<td>2.4 GHz</td>
<td>75 meters</td>
<td>220 Kbps</td>
</tr>
</tbody>
</table>
create new requirements for the enterprise architecture to extend access to applications. However, they do require security measures to make sure the device that is receiving information is a recognized device. It also creates an opportunity for the computing infrastructure to potentially know where a particular device, and most likely the associated user, is located. How these are handled will be discussed later in the description of the proposed service-oriented architecture.

**Wireless LANs**

The set of technical specifications for wireless local area networks (WLANs), labeled 802.11 by IEEE, has led to systems that have exploded in popularity, usability, and affordability. Now wireless LAN can be found in many organizations and public places.

With a wireless LAN, a user’s device is connected to the network through wireless access points (APs). APs are inexpensive—many are available for less than $100—and will usually work perfectly with little or no manual configuration.

Wireless LANs use a standard, called IEEE 802.11, that provides a framework for manufacturers to develop new wireless devices. The first two standards released for wireless LANs were 802.11b and 802.11a. The 802.11b standard was used in most wireless devices in the early adoption of wireless LAN. A new standard, called 802.11g, combines data-transfer rates equal to 802.11a with the range of an 802.11b network. Mimo technology allows the use of two or more antennas operating on that channel. Normally, this would cause interference degradation of the signal because the radio waves would take different paths—called multipath distortion. However, Mimo technology allows the use of two or more antennas operating on that channel. Normally, this would cause interference degradation of the signal because the radio waves would take different paths—called multipath distortion. However, Mimo technology corrects for the multipath effects. IEEE is standardizing the technology as IEEE 802.11n.

For an enterprise, wireless LAN technologies allow pervasive information access throughout the campus. Employees with authorized mobile devices such as wireless laptops and PDAs will be able to get online wherever they are on the campus.

Table 2 summarizes the wireless LAN technologies.
A wireless metropolitan area network (MAN; also referred to as broadband wireless access, or WiMAX) can wirelessly connect business to business within the boundary of a city. It is becoming a cost-effective way to meet escalating business demands for rapid Internet connection and integrated data, voice, and video services.

Wireless MANs can extend existing fixed networks and provide more capacity than cable networks or digital subscriber lines (DSLs). One of the most compelling aspects of the wireless MAN technology is that networks can be created quickly by deploying a small number of fixed-base stations on buildings or poles to create high-capacity wireless access systems.

In the wireless MAN area, IEEE has developed the 802.16 standard (IEEE, 2005b), which was published in April 2002, and has the following features.

- It addresses the “first mile-last mile” connection in wireless metropolitan area networks. It focuses on the efficient use of bandwidth between 10 and 66 GHz.
- It enables interoperability among devices so carriers can use products from multiple vendors. This warrants the availability of lower cost equipment.
- It defines mechanisms that provide for differentiated quality of service (QoS) to support the different needs of different applications. The standard accommodates voice, video, and other data transmissions by using appropriate features.
- It supports adaptive modulation, which effectively balances different data rates and link quality. The modulation method may be adjusted almost instantaneously for optimal data transfer. Adaptive modulation allows efficient use of bandwidth and fits a broader customer base.

The WiMAX technical working group has developed a set of system profiles, standards for protocol-implementation conformance, and test suites (http://www.wimaxforum.org).

One particular technology for WiMAX is non line of sight (NLOS) networking (Shrick, 2002). NLOS networks provide high-speed wireless Internet access to residential and office facilities.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Radio Frequency</th>
<th>Maximum Distance</th>
<th>Data Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>802.11a</td>
<td>5 GHz</td>
<td>20 meters</td>
<td>54 Mbps</td>
</tr>
<tr>
<td>802.11b</td>
<td>2.4 GHz</td>
<td>100 meters</td>
<td>11 Mbps</td>
</tr>
<tr>
<td>802.11g</td>
<td>2.4 GHz</td>
<td>100 meters</td>
<td>54 Mbps</td>
</tr>
<tr>
<td>802.11i</td>
<td></td>
<td></td>
<td>A security standard for encryption on wireless LANs</td>
</tr>
<tr>
<td>802.11n</td>
<td>Varies</td>
<td>Varies</td>
<td>&gt; 100 Mbps</td>
</tr>
<tr>
<td>802.1x</td>
<td>A standard security protocol for user authentication on wireless LANs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NLOS uses self-configuring end points that connect to a PC (personal computer). The end point has small attached antennas and can be mounted anywhere without the need to be oriented like satellite antennas. Two major vendors are Navini Networks and Nokia.

With the wireless MAN technology, enterprises can quickly set up a network to provide wireless access to people in a certain area. It is very useful in situations such as an off-site working session or meeting.

**Wireless NANs**

Wireless neighborhood area networks are community-owned networks that provide wireless broadband Internet access to users in public areas (Schwartz, 2001). To set up a wireless NAN, community group members lend out access to the Internet by linking wireless LAN connections to high-speed digital subscriber lines or cable modems. These wireless LAN connections create network access points that transmit data for up to a 1-kilometer radius. Anyone possessing a laptop or PDA device equipped with a wireless network card can connect to the Internet via one of these community-established access points.

Wireless NANs have been established in more than 25 cities across the United States. Community-based networks differ from mobile ISPs (Internet service providers) such as MobileStar and Wayport that offer subscribers wireless access to the Internet from hotels, airports, and coffee shops. Wireless NANs extend access to consumers in indoor as well as outdoor areas, and the access is typically offered at no charge. For instance, NYC Wireless (http://www.nyc-wireless.net) provides Internet access to outdoor public areas in New York City. In addition, this organization is negotiating with Amtrak to bring wireless Internet access to Penn Station.

Enterprises could leverage the existing wireless NANs and equip employees with the right devices and security mechanisms in order to use these wireless networks to securely connect to the corporate network.

**Wireless WANs**

Wireless wide area networks are commonly known as cellular networks. They refer to the wireless networks used by cell phones.

People characterize the evolution of wireless WAN technology by generation. First generation (1G) started in the late 1970s and was characterized by analog systems. The second generation of wireless technology (2G) started in the 1990s. It is characterized by digital systems with multiple standards and is what most people use today. 2.5G and 3G are expected to be widely available 1 to 3 years from now. 4G is being developed in research labs and is expected to launch as early as 2006.

Wireless WAN originally only offered voice channels. Starting from 2G, people have used modems to transmit data information over the voice network. More recent generations offer both voice and data channels on the same cellular network.

One of the major differentiating factors among the wireless generations is the data transmission speed in which the wireless device can communicate with the Internet. The table below is a comparison of the data transmission rates of the 2G, 2.5G, 3G, and 4G technologies (3Gtoday, 2005). Both 2G and 2.5G include different technologies with different data transmission rates. Global Systems for Mobile Communications (GSM) and Code Division Multiple Access (CDMA) are 2G technologies. General Packet Radio Service (GPRS), CDMA 1x, and Enhanced Data for GSM Environment (EDGE) are 2.5G technologies.

In the United States, cellular carriers Verizon and Sprint use CDMA technology. Cingular uses GSM, GPRS, and EDGE technologies. Both Verizon and Sprint have rolled out their CDMA 1x
services, which is 2.5G. Cingular has rolled out GRPS service and is starting to roll out EDGE service in selected markets.

Wireless WANs are available wherever cell phones can be used. For now, they are the most pervasive wireless networks. By subscribing to a service plan, an enterprise user’s laptop computer or other mobile device can connect to the Internet through the service provider’s cellular towers.

**Ultrawideband (UWB)**

Traditional radio-frequency technologies send and receive information on particular frequencies, usually licensed from the government. Ultrawideband technology sends signals across the entire radio spectrum in a series of rapid bursts.

Ultrawideband wireless technology can transmit data at over 50 Mbps. A handheld device using this technology consumes 0.05 milliwatts of power as compared to hundreds of milliwatts for today’s cell phones. Ultrawideband signals appear to be background noise for receivers of other radio signals. Therefore it does not interfere with other radio signals. Ultrawideband is ideal for delivering very high-speed wireless-network data exchange rates (up to 800 Mbps) across relatively short distances (less than 10 meters) with a low-power source.

Another feature of ultrawideband signals is that they can penetrate walls. Therefore, this technology would allow a wireless device to communicate with a receiver in a different room. This feature can also be used to detect buried bodies, people in a building, or metal objects in concrete.

**Mesh Radio and Mess Networks**

Mesh radio is a wireless network technology that operates in the 28-GHz range of the radio spectrum and provides high-speed, high-bandwidth connectivity to the Internet (Fox, 2001). A mesh radio network consists of antennas connected in a web-like pattern to a fiber-optic backbone. A single antenna attached to the roof of a building could provide Internet access to all of the subscribers residing in the building. Each node on the network has a small, low-power, directional antenna that is capable of routing traffic for other nodes within a 2.8-kilometer radius. In contrast to other wireless networks, mesh radio avoids many of the line-of-sight issues between the base station and each node on the network. Consequently, the configuration of mesh radio reduces the chance of encountering physical obstructions that could impede access to the network.

Mesh radio networks are being developed in two different ways. CALY Networks has devel-
Leveraging Pervasive and Ubiquitous Service Computing

oped a system that utilizes the Internet protocol (IP) as its communication mechanism, while Radiant Networks has created a system that communicates using the asynchronous transfer mode (ATM). Providers of mesh radio services include British Telecommunications, TradeWinds Communications (http://www.tnsconnects.com), and Nsight Teleservices (http://www.nsighttel.com).

Features of mesh radio include the following:

- Provides upload and download data rates of up to 25 Mbps
- Supports up to 600 subscribers per square kilometer without degradation of service
- Provides cost-effective access to broadband services in rural communities or urban areas
- Increases network capacity and resilience as the customer base grows

Different from the mesh radio technology, mesh networks enable wireless devices to work as a peer-to-peer network, using the handsets themselves instead of the radio towers to transmit data (Blackwell, 2002). Each handset would be capable of transmitting data at rates from 6 Mbps to 18 Mbps. This technology can be used for a group of users or devices communicating with each other in a peer-to-peer mode without needing an established wireless network. The technology was developed by Mesh Networks Inc., which has been acquired by Motorola.

Sensor Networks

Motes (also called sensor networks or Smart Dusts; Culler & Hong, 2004) are small sensing and communication devices. They can be used as wireless sensors replacing smoke detectors, thermostats, lighting-level controls, personal-entry switches, and so forth. Motes are built using currently available technology and are inexpensive enough to be deployed in mass quantities. Depending on the sensors and the capacity of the power supply, a mote can be as big as 8 cubic centimeters (the size of a matchbox) or as small as one cubic millimeter.

Motes are the result of a joint effort between Defense Advanced Research Projects Agency (DARPA) and the University of California, Berkeley, research labs. Most initial applications are positioned to helping the military for tasks such as surveillance of war zones, the monitoring of transportation, and the detection of missiles and/or biological weapons. Commercial mote sensors are available from Crossbow Technology.

A mote is typically made up of the following:

- A scanner that can scan and measure information on temperature, light intensity, vibrations, velocity, or pressure changes
- A microcontroller that determines tasks performed by the mote and controls power across the mote to conserve energy
- A power supply that can be small solar cells or large off-the-shelf batteries
- TinyOS, an open-source software platform for the motes. TinyOS enables motes to self-organize themselves into wireless network sensors.
- TinyDB, a small database that stores the information on a mote. With the help of TinyOS, the mote can process the data and send filtered information to a receiver.

These motes enable enterprises to constantly collect important information and send the information to the appropriate server for processing so that the appropriate response can be initiated when necessary. The motes become the generator of pervasive information that reflects the status of business processes or environmental conditions.
Pervasive Devices

Pervasive devices come in different forms and shapes. Compared to a networked computer, some pervasive devices, such as landline or cell phones, are more widely available. Other devices are simply more portable and thus can be easily carried around. Yet other devices are embedded in the environment and are able to deliver specialized information. In terms of their functions, some are for accessing the Internet, some are just for entering information while the user is on the go, and others are for storing large amounts of information and can be easily carried around.

Traditional Telephones, Pagers, and Cell Phones

Traditional landline telephone has been the most pervasive communication device around the world. Voice markup languages such as VoiceXML (voice extensible markup language; Rubio, 2004), together with supporting technologies such as the voice browser and voice gateway, has made the traditional telephone yet another device for connecting the user to the Internet. With speech recognition, users can choose to use touch tone or simply say what they need. Figure 1 shows how a telephone can be used to connect to the Internet.

1. The user dials the number from any phone (landline or mobile).
2. The call is routed to the corresponding voice gateway, which maps the phone number to a particular application hosted at the enterprise network.
3. The voice gateway knows the URL (uniform resource locator) of the application. It uses an HTTP (hypertext transfer protocol) request to fetch the first dialog of the application.
4. The enterprise Web server and application server return the dialog to the gateway in the form of a VoiceXML document.
5. The gateway interprets the VoiceXML document, plays the greeting, and asks the user for input. Now the user can use touch tone or speech to provide input. Based on the user input and the application logic as described in the VoiceXML file, the voice gateway decides what dialog to fetch next from the enterprise network.

Pagers allow users to receive alerts with a limited amount of text. With two-way pagers, users can also reply with a text message.

With cell phones (not smart phones), besides the same communication capabilities of a landline telephone, most users can use short-message service (SMS) to send and receive text messages. This is good for near-real-time conversational communications.

Figure 1. Voice gateway connects the phone network with the data network
Smart Phones, Wireless PDAs, and Blackberry Devices

Smart phones are cell phones that have both voice and data capabilities. Such a cell phone comes with a mini Web browser and thus can be used to access Internet content. However, since the smart phones typically have rather small screens, they can only access pages specifically designed for small screens and coded in a special markup language such as Wireless Markup Language (WML). Some smart phones are equipped with a computing platform such as the Java Virtual Machine that can run applications written in J2ME (Java 2 Micro Edition).

Wireless PDAs typically have larger screens than cell phones and can directly access HTML (hypertext markup language) pages. Some wireless PDAs can also be used to make phone calls, and are referred to as PDA phones. Since many people prefer to carry only one of such mobile device around, there is a competition between PDA phones and smart phones, a war in which the smart phones seem to be winning.

ViewSonic (http://www.viewsonic.com) made a super-sized PDA, called the ViewPad, that offers a regular 800x600-pixel screen. The ViewPad can be a very useful mobile device when regular screen size is a necessity while light weight and zero-boot-up time are also desired.

Blackberry devices made by Research in Motion (http://www.rim.com) has been a big success for enterprise users as they provide a very convenient way for reading and typing e-mails while being away from the office.

Laptop or Tablet PCs with Wireless Access

When size and weight are not inhibitive, mobile users may choose to carry a laptop or tablet PC while on the go. These mobile PCs use wireless cards to connect to either a wireless LAN or wireless WAN. Many such laptops now have built-in wireless LAN cards, and have slots for users to insert a wireless WAN card such as the AirCard made by Sierra Wireless (http://www.sierrawireless.com). An enterprise also needs to be prepared to provide support to mobile users in order to help them connect to the Internet through Wi-Fi hot spots (Hamblen, 2005).

Wireless LAN is often available at a corporation campus, or at public hot spots such as many airports and Starbucks Cafés. Wireless WAN is available wherever cellular service is available for the specific provider that the wireless card is registered with.

IP Phones

IP phones are telephones that use a TCP/IP (transmission-control protocol/Internet protocol) network for transmitting voice information. Since IP phones are attached to the data network, makers of such devices often make the screens larger so that the phones can also be used to access data. What makes IP phones pervasive devices is that a user who is away from his or her own desk can come to any IP phone on the same corporate network, log in to the phone, and make the phone work as his or her own phone. The reason is for this is that an IP phone is identified on the network by an IP address. The mapping between a telephone number and an IP address can be easily changed to make the phone “belong” to a different user.

In terms of the information-access capability, Cisco (http://www.cisco.com) makes IP phones that can access information encoded in a special XML format. Example applications on the phone include retrieving stock quotes, flight departure and arrival information, news, and so forth.

Pingtel (http://www.pingtel.com) developed a phone that runs a Java Virtual Machine. This makes the phone almost as powerful as a computer.

Mitel (http://www.mitel.com) made an IP phone that allows a user to dock a PDA. With this capability, users can go to any such IP phone,
Leveraging Pervasive and Ubiquitous Service Computing

dock their PDA into the phone, and immediately have their address books on the PDA available to the telephone. Users can also have their personal preferences transferred from the PDA to the phone and start to use the phone the way they prefer. In addition, users can benefit from new applications on the PDA, such as portable voice mail and dialing by the address book.

The Mitel IP phone seamlessly blends the wired and wireless world for the user so that they are no longer dealing with two separate communication tools. It also provides users with location transparency within the network.

Orbs (Ambient Devices)

Orbs are simple devices that convey information at a glance in a manner that is easy to observe and comprehend (Feder, 2003). Orbs only present a visual indication of the data, not detailed information or actual numbers. Orbs come in different forms. One common orb is a simple globe that changes color and intensity. Other forms include the following:

- Wall panels that adjust color or blink
- Pens, watch bezels, and fobs that change color
- Water tubes that vary the bubble rate
- Pinwheels that change speed

Orbs operate via wireless pager networks under the command of a server. This server gathers pertinent information from sources, including the Web, condenses it to a simple value, and periodically sends the information to the orbs.

Orbs are currently available from several retailers. The wireless service costs about $5 per month per device. Ambient Devices (http://www.ambientdevices.com) sells orbs and provides the communications service.

The information displayed by orbs is configurable. There are currently available data feeds for stock-market movement and weather forecasts.

Input Technologies: Dictation, Anoto Pen, and Projection Keyboard

Two natural ways for mobile users to input information are speech and handwriting.

Speech input can be at two levels: question-and-answer vs. dictation. Question-and-answer speech input is useful for entering structured information where the answers can be predefined using a grammar. Dictation technology allows users to speak freely and tries to recognize what the user has said. Diction technology typically requires a training phase to tune the speech recognizer to each particular speaker in order to achieve high recognition accuracy. Leading dictation products are Dragon NaturallySpeaking from ScanSoft (http://www.scansoft.com) and ViaVoice from IBM (http://www.ibm.com).

The Swedish company Anoto (http://www.anoto.com) invented a technology for pen-based input (McCarthy, 2000). It consists of a digital pen that feels like a regular ballpoint pen, a special paper with patterns of dots printed on it, and a wireless technology such as Bluetooth that sends handwritten information stored in the pen to a computer. As the user writes, the pen not only records what has been written, but also the order in which the user writes it. Anoto has partnered with companies such as Logitech (http://www.logitech.com) and Nokia (http://www.nokia.com) to bring this technology to end users.

For users who want to use a keyboard without carrying one, Canesta (http://www.canesta.com) developed the projection keyboard, in which the image of a keyboard is projected on a surface. By typing on the projection keyboard, information is entered into the associated PDA device.

Application Scenarios

From an enterprise’s perspective, the following applications areas are where pervasive computing brings business value.
Leveraging Pervasive and Ubiquitous Service Computing

- Allow employees to stay in touch with phone calls, voice mail, e-mail, and so forth while being away from the office.
- Give employees access to information or transactions via mobile devices while on the road.
- Provide employees with access to the corporate network from anywhere on the Internet (i.e., remote access).
- Send location-based information to employees and customers.
- Monitor device status, perimeter security, and so forth using a wireless sensor network.

COMMUNICATION: UNIFIED COMMUNICATION AND INSTANT COMMUNICATION

With cell phones and pagers, it is not very hard to keep mobile users in touch. But some pervasive communication technologies have reached a higher level. Let us look at two such technologies: unified communication and instant communication.

Unified communications refers to technologies that allow users access to all their phone calls, voice mails, e-mails, faxes, and instant messages as long as they have access to either a phone or a computer. With a computer, a software phone allows the user to make or receive phone calls. Voice-mail messages can be forwarded to the e-mail box as audio files and played on the computer. Fax can be delivered to the e-mail box as images. With a phone, a user can listen to e-mail messages that the system would read using the text-to-speech technology. A user can request a fax to be forwarded to a nearby fax machine.

Unified communications services are offered by most traditional telecommunications technology providers such as Cisco, Avaya, and Nortel.

Instant communication refers to the ability of reaching someone instantly via a wearable communication device. Vocera (http://www.vocera.com) offers a system that uses 802.11b wireless local area networks to allow mobile users to instantly communicate with one another. Each user only needs to have a small wearable device to stay connected. To reach someone, the user would only need to speak a name, a job role, a group, or a location to the system, and the system will take care of the rest. By combining a small wearable device and the speech-recognition capability, Vocera offers a highly usable solution for mobile communication within an organization.

The functions and features of Vocera include the following:

- Instant communication via a small wearable device and speech commands
- Hands-free communication. Except for pressing the button to start and stop a conversation, a user’s hands are free during the communication.
- Flexibility in how to specify the recipients. A user can use a name, role, group, or location to tell the system whom to contact.
- The option of having a conversation or leaving a message, for both one-to-one and group communications
- Call controls such as call transfer, blocking, or screening
- Outside calling through the private branch exchange (PBX). The Vocera server can be connected to the PBX to allow users of Vocera to contact people outside the organization.

The Vocera technology has been well received in organizations such as hospitals where users’ hands are often busy when they need to communicate with others.
Leveraging Pervasive and Ubiquitous Service Computing

Mobile Access to Information and Applications

Organizations can benefit significantly by allowing mobile access to information and applications. Here are a few examples.

Sales-Force Automation

Salespeople are often on the road. It is important for them to have access to critical business information anywhere at anytime. Pervasive access to information increases their productivity by using their downtime during travel to review information about clients and prospects, about the new products and services they are going to sell, or to recap what has just happened during a sales event when everything is still fresh in their memory. Being able to use smart phones or wireless PDAs to conduct these activities is much more convenient for salespeople as opposed to having to carry a laptop PC.

Dashboard or Project-Portfolio Management

For busy executives, it is very valuable for them to be able to keep up to date on the dashboard while they are away from the office and to take actions when necessary. It is also very helpful for them to be able to look at the portfolio of projects they are watching, update information they have just received during a meeting or conversation, and take notes or actions about a specific project.

Facility Management and Other On-Site Service Applications

Mobile access to information can significantly boost the productivity of on-site service people such as facility- or PC-support staff. With mobile access, they can retrieve ticket information on the spot, update the ticket as soon as they are done with the work, and get the next work order without having to come back to the office. Mobile access also reduces the amount of bookkeeping, which requires a lot of manual intervention, and thus reduces the chance of human errors.

Remote Access to Corporate Network

Allowing employees access to the corporate network from anywhere on the Internet could certainly bring convenience to employees and boost productivity. There are two primary fashions of allowing remote access.

One approach is through a technology called virtual private network, or VPN. This typically requires the user to carry a laptop offered by the employer. Once the user is connected to the Internet, a secure connection (called a VPN tunnel) is established between the laptop and the corporate network after both user and device are authenticated. Then the user will have access to all the information and applications just as if the user were in the office.

The other approach does not require the user to carry a corporate laptop. It simply requires that the user has access to a Web browser. In this case, for security reasons, two-factor authentication is often employed, in which the user not only needs to provide a user ID and password, but also something else, such as the security code generated by a hard token. With this approach, an enterprise can choose which applications to make available for remote access. Terminal service technology offered by Citrix (http://www.citrix.com) can be used to offer browser-based remote access to applications, both Web based and desktop based.

Location-Based Services

A special type of pervasive application is location-based service. With wireless LANs, when a mobile user is in the vicinity of an access point,
location-specific information can be delivered to
the user’s mobile device. With wireless WANs, a
user’s location can be determined by the cellular
tower(s) that the user’s handset is communicating
with, or by the GPS (Global Positioning System)
receiver the user is using. Location-based services
include pushing information about local busi-
nesses, sending promotions to the user’s device
based on the user’s profile and preferences, and
showing meeting agendas and meeting material
if the user is on the meeting attendee list for the
room at the time.

If location needs to be accurately determined,
an ultrasonic location system called the bat system
can be used. This 3-D location system uses low
power and wireless technology that is relatively
inexpensive. An ultrasonic location system is
based on the principle of trilateration: position
finding by the measurement of distances. A short
pulse of ultrasound is emitted from a transmitter
or bat that is attached to a person or object to be
located. On the ceiling are receivers mounted at
known points. These receivers can measure the
pulse and length of travel.

An ultrasonic location system is composed of
three main components.

• **Bats:** Small ultrasonic transmitters worn by
  an individual or on an object to be located
• **Receivers:** Ultrasonic signal detectors
  mounted in the ceiling
• **Central controller:** Coordinator of the bats
  and receivers

To locate a bat, the central controller will
send the bat’s ID via a 433-MHz bidirectional
radio signal. The bat will detect its ID through
the embedded receiver and transmit an ultrasonic
signal containing a 48-bit code to the receiver in
the ceiling. The central controller will measure
the elapsed time that it took for the pulse to reach
the receiver. The system developed at the AT&T
Cambridge facility can provide an accuracy of 3
centimeters.

Overall, location-based services are still in
research mode. Once the technology becomes
mature and “killer apps” are identified, there
could be an explosive adoption.

**User-Interaction Models**

In the context of pervasive computing, it is usually
inconvenient, if not impossible, for the user to
enter text using a regular keyboard. Sometimes,
it is also inconvenient for the user to read text.
Therefore, other input and output mechanisms
have to be employed.

Nontraditional input mechanisms include
speech recognition, gesture, touch screen, eye
gazing, software keyboard, and projection key-
board. Among these, a combination of speech-
recognition and pen-based touch-screen input is
most natural for most situations. This is also what
PDAs and tablet PCs typically offer.

Nontraditional output mechanisms include
converting text to speech and using sound,
blinking, and vibration to convey information
(as in ambient computing described earlier in
this chapter).

Multimodal interaction allows a user to choose
among different modes of input and output. For
mobile users, speech is typically the most conve-
nient way for input, while visual means may still
be the most powerful way of seeing the output
(especially when the output includes pictures or
diagrams).

Kirusa (http://www.kirusa.com) has developed
technologies to support multiple levels of multi-
modal interaction. SMS multimodality allows users
to ask a question in voice and have the answers
delivered to their mobile devices in the form of
an SMS message. Sequential multimodality al-
lows users to use the interaction mode deemed
most appropriate for each step of the process.
Simultaneous multimodality lets users combine
different input and output modes at the same time.
For example, for driving directions, a user can
Leveraging Pervasive and Ubiquitous Service Computing

say “I need directions from here to there,” while pointing to the start and end points.

Both IBM and Microsoft have developed technologies that will support multimodal interaction. IBM’s solution is based on the XHTML+VoiceXML (or simply X+V) specification. Microsoft’s solution is based on the speech application language tags (SALT) specification, which defines speech input and output tags that can be inserted into traditional Web pages using XHTML.

Besides deciding on what interaction mode to support, much effort is needed to apply user-centered design in order to deliver a good use experience for mobile users (Holtzblatt, 2005).

A Service-Oriented Architecture to Support Pervasive Computing

For an enterprise to leverage pervasive computing, instead of deploying various point solutions, the better way is to build an architecture that is well positioned to support pervasive devices and usage. In order to provide mobile users with maximum access to enterprise information and applications with customized interaction methods and workflow, and at the same time minimize the extra cost in supporting pervasive access, a service-oriented architecture should be established.

The following picture shows a service-oriented architecture that supports pervasive computing. Let us look at this architecture from the top to the bottom.

• Users access applications from different devices. Some devices, such as the regular telephone, have only the voice channel. Some, such as the Blackberry devices, only have the visual display. Others may have both voice and visual channels. The size of the visual display ranges from 1 inch for cell phones, several inches for the PDA and Blackberry, and 15 or more inches for laptops.

• The user devices may be on different network connections, ranging from wireless LAN and wireless WAN to telephone networks.

• Users access applications through the applications’ end points. This could be a URL, a phone number, or a start screen stored on the end user’s device.

• The pervasive layer sits between the application end points and the SOA layer to provide services to specifically support the mobile users.

• The device identification engine uses a unique ID to identify the device the user is using. This requires the augmentation of some of the communication protocols to include a universally unique ID (such as the radio frequency identifier, or RFID) of the device that is initiating the request. With this ID, the system can uniquely identify the device and thus have knowledge of its capabilities, the associated user, and so forth. The ID information is also passed to the security service in the SOA layer to help decide whether the user is authorized to access the application.

• The access-control engine uses information about the device and the communication channel it is coming from to determine the best way to communicate with the device: voice only, visual only, SMS, or some type of multimodal interaction.

• Based on the desired interaction mode with the user, the content-transformation engine either calls the appropriate version of the application or dynamically transforms the information into the appropriate markup language: HTML, WML, VoiceXML, X+V, and so forth, using the eXtensible Stylesheet Language transformation (XSLT) technology.
• The location-determination service uses mechanisms built into the networks to determine the geographic location of the user, and then decides whether the information should be tailored based on the location and whether additional location-based information should be pushed to the user.
• The session-persistence engine uses the device ID and user-identity information to keep the user in the same session while the user is roaming from one network to another, or from disconnected mode to connected mode again during a short period of time. For smart-client applications, where data may be temporarily stored on the device when connection is lost, the session-persistence layer would also take care of synchronizing the data on the device with data on the server.

Figure 2. A service-oriented architecture that supports pervasive computing
• The business-application composition layer uses information received from the pervasive layer to determine how to integrate the business services together to best fit the need of this mobile user.

• The SOA layer provides the business services and technical services that are integrated together through the enterprise service bus. The business services can be built using Net-centric technologies such as Java or Microsoft .NET, or they can be built based on existing legacy business functions such as customer information control system (CICS) transaction and stored procedures. They can also be based on business functions built using third-party tools or existing in-house business modules developed in C or C++, COBOL, and so forth.

After the establishment of such an architecture (which can be gradually built across multiple projects), when building an application that supports pervasive access, business services are either reused or built. When an existing business service is reused, the project team needs to go through the service’s specification to verify that the service will work well with the pervasive layer to support all the pervasive channels that the application is going to support. If not, the first thing the project team should try to modify is the pervasive layer. If the issue really lays in the fact that the business service is not a “pure” service, that is, the service is tied to access methods, then the service needs to be either modified or wrapped in order to support the new requirements. When such a modification occurs, the service needs to be made backward compatible, and existing applications that use the service need to be regression tested. Eventually, the service definition needs to be modified to reflect the changes.

With this architecture, when a new pervasive access channel appears, or there is a change in an existing channel, then the only thing that needs to be modified is the pervasive channel. All business services can remain the same.

FUTURE DIRECTIONS

Moving forward, there needs to be much research and development work on building a system infrastructure that can use different sources of information to judge where the user is, and what devices and interaction modes are available to the user during a pervasive session. This will enable smarter location-based information push to better serve the user.

A related research topic is how to smoothly transition an interaction to a new device and interaction mode as the user changes locations and devices. Some initial work on this subject, referred to as seamless mobility, is being conducted at IBM and other organizations.

Another area that deserves much attention is the proactive delivery of information that users will need based on their profiles and information such as activities on their calendars or to-do lists. This relates to previous research efforts on intelligent personal assistants with integration into the pervasive computing environment.

REFERENCES


*This work was previously published in Enterprise Service Computing: From Concept to Deployment, edited by R. Qiu, pp. 261-284, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 1.21
Service Provision for Pervasive Computing Environments

Emerson Loureiro
Federal University of Campina Grande, Brazil

Frederico Bublitz
Federal University of Campina Grande, Brazil

Loreno Oliveira
Federal University of Campina Grande, Brazil

Nadja Barbosa
Federal University of Campina Grande, Brazil

Angelo Perkusich
Federal University of Campina Grande, Brazil

Hyggo Almeida
Federal University of Campina Grande, Brazil

Glauber Ferreira
Federal University of Campina Grande, Brazil

INTRODUCTION

The fast development on microelectronics has promoted the increase on the computational power of hardware components. On the other hand, we are facing a significant improvement on energy consumption as well as the reduction of the physical size of such components. These improvements and the emergence of wireless networking technologies are enabling the development of small and powered mobile devices. Due to this scenario, the so-called pervasive computing paradigm, in-
Service Provision for Pervasive Computing Environments

introduced by Mark Weiser in 1991 (Weiser, 1991) is becoming a reality. Such a paradigm envisions a world where environments are inhabited by computing devices, all of them seamlessly integrated into peoples’ lives, and effectively helping to carry on their daily tasks.

Among others, one major characteristic of Weiser’s vision is that each device in an environment becomes a potential client or provider of resources. Not surprisingly, pervasive computing environments are becoming dynamic repositories of computational resources, all of them available to mobile users from the palm of their hands. However, devices can unpredictably join and leave such environments. Thus, resources can be dynamically made available or unavailable. Such a scenario has a great impact on the way that resources are found and used. In the case of static environments, such as the Web, it is reasonable to look up and access resources, such as Web pages, knowing the address of their providers beforehand. On the other hand, for dynamic environments, such as the pervasive computing ones, this is not a reasonable approach. This is due to the fact that one cannot guarantee that the provider of a resource will be available at any moment, because it may have left the environment or simply turned off. A better approach would be to discover these resources based on their descriptions, or any other feature that does not require the client to know the specific address of their providers.

To this end, some of the current pervasive computing solutions, like Wings (Loureiro, Bublitz, Oliveira, Barbosa, Perkusich, Almeida, & Ferreira, 2006), Green (Sivaharan, Blair, & Coulson, 2005), RUNES (Costa, Coulson, Mascolo, Picco, & Zachariadis, 2005), and Scooby (Robinson, Wakeman, & Owen, 2004), are making use of a novel approach from the branch of distributed applications, the service-oriented computing paradigm (Papazoglou, 2003; Huhns & Singh, 2005). This is due to the fact that such a paradigm provides a crucial element for pervasive computing systems, the ability for dynamically binding to remote resources (Bellur & Narendra, 2005), which enables mobile devices to find needed services on demand.

However, pervasive environments may be structured in different ways. They can range from wired networks to completely wireless ones, where communication among the devices is performed in an ad hoc way. Such a characteristic indicates that the way services are provisioned in a pervasive computing environment should fit in its organization, in order to enhance the access to the services available.

Considering the above discussion, in this article we provide a review on service provision and its applicability in pervasive computing. More precisely, we will list the existing service provision approaches and discuss the characteristics and problems associated with each one, as well as their usage in pervasive computing environments. We start by providing introductory concepts of service-oriented and pervasive computing, respectively in the service-oriented computing and pervasive computing sections. Next, we present the service provision techniques available and how they can be applied for pervasive computing environments. The main current solutions within this scope will be introduced in the service oriented technologies section. Some of the future trends associated with research for service provision in pervasive computing environments will be presented in the future research trends section. Finally, in the conclusions section we present the conclusions of this article.

**SERVICE-ORIENTED COMPUTING**

The service-oriented computing (SOC) paradigm has been considered as the next step in distributed computing (Papazoglou, 2003). In a general way, this paradigm can be viewed as the development of applications through the runtime integration of software pieces named of services (McGovern, Tyagi, Stevens, & Mathew, 2003). In this process,
three elements are involved: defining what is known as a service-oriented architecture (SOA), a service client, a service provider, and a service registry. The former is the one who wishes to use a service. Conversely, service providers are those which offer services for potential clients. Finally, the service registry is where providers advertise or announce their services (through service advertisements), enabling clients to dynamically discover them. By dynamic, we mean that clients are capable of discovering services at runtime, thus providing a high degree of flexibility for applications. Once clients have discovered a service, they are able to bind to it; that is, to create a link with the service, in order to use it (through a proxy to the real service). This process of advertising, discovering, binding, and using a service is illustrated in Figure 1.

In open environments, the dynamic discovery of services implies that they can be used by heterogeneous clients. Within this scope, heterogeneity is concerned with features like the operating system running in each client and the hardware platform it has been built on. As a consequence of such heterogeneity, for enabling an application to flexibly integrate services, they should present the following features (Papazoglou, 2003):

- **Loose Coupling**: A service must not require from the clients any knowledge about its internal implementation.
- **Implementation Neutrality**: The usage of services must not rely on any specific programming language, operating system, or hardware platform.
- **Dynamically Discoverable**: Services should be discovered at runtime.

**PERVASIVE COMPUTING**

The field of pervasive computing has its origins at the Xerox Palo Alto Research Center. The pioneer work that has been led there has culminated in the novel article of Mark Weiser in 1991 (Weiser, 1991), where he describes the first ideas of pervasive computing. Weiser’s vision is at the same time revolutionary and simple: a world where computing is embedded in everyday objects, like cars, televisions, and air conditionings, all seamlessly integrated into our lives and performing tasks for us (Turban, Rainer, & Potter, 2005). When Weiser talked about seamless integration, he meant that applications running in these objects should act proactively on our behalf. They should, for
example, present us with relevant information, based on what we want/need and the resources (e.g., a printer) available in the environment we are immersed.

**SERVICE PROVISION APPROACHES IN PERVERSIVE COMPUTING**

When provisioning services in a pervasive environment, one aspect to be considered is the way it is organized; that is, whether the environment is based on a wired network infrastructure, whether it is formed in an *ad hoc* way, or both. This is necessary for dealing with the particularities of each environment, and within this scope, we can say that there are two major ways of performing service provision (Nickull, 2005): the *push-based* and the *pull-based* approach. In the next sections we will outline each of these approaches as well as describe how they fit into pervasive computing environments.

**Push-Based Service Provision**

In this approach, a provider advertises its services directly to potential clients. In other words, it sends service advertisements to all network hosts, whether they are interested on the service or not. Such advertisements, when received by a host, will be kept in a local service registry. Therefore, once a client wants to discover some service, it will then look up in such a registry for the services that match its needs. An example of the push-based service provision is illustrated in Figure 2. Note that the provider, host A, sends the advertisement of a service directly to hosts B and C, as illustrated in Step 1. When this advertisement is received, it will be stored on a local service registry at each host (Step 2). Then, once a host, in our example host C, wants to discover a service, it then inquires this registry, as illustrated in Step 3. Finally, considering that a relevant service has been found, it is possible to ask its provider (host A) for a proxy to the service (Step 4), enabling host C to use it.

Using this approach, one major problem to be pointed out is about the validity of the service advertisements. It is concerned with the fact that a service provider can leave the network, but the advertisements associated with its services can still be available. One approach which could be used for solving this problem is to require the provider to explicitly notify the network hosts about its leaving, and consequently its services will be no longer available. The problem is that it is not always possible to do that. For example, if the provider is a mobile device, it may be suddenly run out of energy. To deal with this, providers could be aware of the energy level of the device, in order to notify the network hosts that within some minutes it may not be accessible anymore. However, other factors can be involved in the leaving of a provider from the network. It can be just turned off by its user, or leave the network coverage area. Keeping track of all these factors is a task that certainly overloads the provider. A more reasonable solution would be requiring both providers and clients to cooperate for renewing the service advertisement. Therefore, as long as the advertisements are renewed, it is possible, but not guaranteed, that the service is available. On the other hand, when the advertisement has not been renewed within a time interval, then the service is probable, but also not guaranteed, to be unavailable, either because its provider has left the network or because the service has been unadvertised.

One interesting point to notice is that the push-based service provision does not require any physical infrastructure. This means that such an approach is well suited in decentralized and/or infrastructure-less pervasive environments. One problem, in the scope of pervasive computing, is that the advertisement task can consume a lot of bandwidth if many devices are provisioning services in the environment. Therefore, in envi-
environments with very limited wireless links this is certainly a major problem. On the other hand, as services are searched locally, the discovery process does not involve costs of communication.

Pull-Based Service Provision

In the pull-based approach, in order to discover services clients must inquiry remote registries for the needed services. This can be performed in two ways; either using centralized or distributed registries.

Centralized Provision

The centralized service provision consists in scattering service registries in specific servers (i.e., registry servers) of the network. Therefore, for advertising a service, the provider must initially find which of these servers are available in the network. After that, it has to determine in which of them the service will be advertised (instead, the provider could advertise the service in all the available servers). When a client wants to discover services, it must also find the registry servers available in the network, and then discover the services advertised in them. It is important to notice that, once the registry servers are found, unless they become unreachable, clients and providers do not need to discover them anymore. Jini and Web Services are some of the current technologies that support centralized service provision. In Figure 3 we illustrate an example of such an approach. In such a figure, services are advertised by hosts A and B (Step 1), in a single registry server, host C (we are considering that the clients have already found the registry server). After that, each advertisement is stored in the service registry maintained by host C (Step 2). Also considering that host D has already found

Figure 2. Example of a push-based service provision
the registry server, it is then able to inquiry such server for the advertised services (Step 3). When a relevant service is found (Step 4), host D can interact with its provider, in this case host A, to retrieve the proxy for the service (Step 5).

A problem with this approach is that, if all these servers are off the network, services can not be discovered, even when their providers are available. In these cases a possible solution could be the election of new servers from the moment that is detected that the others are no longer available. Furthermore, the centralized service provision raises the same problem of the pull-based one concerned with the validity of the service advertisements. In this way, the same, or at least similar, solutions can be applied here.

As the centralized service provision requires the existence of registry servers, it is not well suited for highly dynamic pervasive environments. In these cases, as nodes join and leave the environment all the time, there would be too many changes in the current registry servers, which would in turn degrade the provisioning of services. On the other hand, this approach is very useful for environments equipped with wired network. In such environments, services can be deployed in the wired network and thus be accessed by mobile clients through wireless links. In environments populated with lots of mobile clients, this is certainly a good choice, as the bandwidth available in the wired network could support a great number of accesses.
Distributed Provision

In the distributed service provision, services are advertised in registries located in each host. Undoubtedly, in this approach the advertising task is easier to be performed than in the other ones, as it does not involve sending advertisements to central servers or directly to the other hosts. However, service discovery is more complicated, as it must be performed by inquiring each available host for the needed services. As no centralizer hosts are necessary for advertising services, discovery is possible whenever a client and a provider are present in the network. An example of the distributed service provision approach is illustrated in Figure 4. Initially, each host advertises its services (Step 1). Once a client needs to perform service discovery, in our example host A, it asks every host in the network for the needed service (Step 2). It is important to note the possibility of redirecting the service discovery request to hosts that are not reachable from the client. In Figure 4 this is performed by host B when it redirects the request of host A to C. When a host has a service matching the client’s needs, it sends a notification (Step 3). From this notification, the client can then retrieve a proxy to the service (Step 4).

The major problem with this approach is associated with the protocols for performing service discovery. As any node in the network is a potential service provider, the discovery protocol must be well designed, in order to cover all hosts of the network. If any host is missing in the discovery process, it is possible that a relevant service may not be found. In \textit{ad hoc} pervasive environments, one possible solution to this problem is first to dis-

\textbf{Figure 4. Example of a distributed service provision}
cover the hosts in the vicinity, and then to retrieve
the relevant services they provide. However this
solution only performs well in small networks,
where the available hosts can be discovered from
any other one. Another point to be considered is
about the network congestion that such service
discovery protocols can cause. As the search
should include all hosts in the network, the protocol
must apply techniques for avoiding flooding the
network. Obviously, this is not a serious problem
in wired networks, but considering the wireless
ones, it must be strictly taken into account. A
good usage scenario of the distributed provision
approach is a decentralized pervasive environ-
ment where the edge of the network is formed by
mobile clients and its core is populated by service
providers connected through a wired network.

SERVICES ORIENTED
TECHNOLOGIES

In this section we present the main technologies
related to the provision of services in pervasive
computing environments.

Universal Plug and Play (UPnP)

The Universal Plug and Play (Richard, 2000)
is an open architecture, which uses standard
Internet protocols for pervasive peer-to-peer
network connectivity (http://www.upnp.org). The
UPnP protocol defines a set of steps, addressing,
discovery, description, control, eventing, and
presenting, which enables the automatic network
configuration and service provision. Through the
addressing, devices get their network address,
which is performed by a dynamic host con-
figuration protocol (DHCP). The discovery step
consists of notifying the other hosts, through the
push-based approach, about the services, and em-
bedded devices, that a joining host provides. The
discovery can also be performed in a distributed
cap-based fashion. The next step, description,
is about the description of a device, stored as an
XML document. In such a document, it is kept,
among other information, the description of the
services that the device provides. Therefore, by
discovering a device, it is possible to retrieve the
services it provides, and then, through the control
step, invoke their actions. Changes in a service
can be notified to interested devices through the
eventing step. Finally, through the presenting step,
it is possible to load a specific URL, specified in
the device description, in order to display a user
interface for controlling a device.

Jini

Jini is a service-oriented Java technology based on
a centralized pull-based approach (Waldo, 1999).
Therefore, service advertisements are stored in
central servers, which are named lookup servers.
Jini uses the RMI (http://java.sun.com/products/
jdk/rmi - Remote Method Invocation) protocol
for all interactions involved in the advertisement,
discovery, and invocation of services. When a
client discovers and binds to a service, it is in-
corporated to the client by downloading the code
of a proxy to the required service, named remote
control object.

The Jini platform uses the concept of lease
for controlling the access to the services. A lease
is a sort of warrant that a client has for using a
service during a specific period of time. When
the lease expires the client needs to renew it
with the provider if it wishes to continue using
the service.

Bluetooth

Bluetooth is a standard for wireless communica-
tion among small devices within short distances
(Johansson, Kazantzidis, Kapoor, & Gerla, 2001),
defining higher-level protocols for both host and
service discovery (http://www.bluetooth.org). The
discovery of services in the Bluetooth standard is
defined by the service discovery protocol (SDP),
which enables to enumerate the devices in the vicinity and retrieve the services they provide.

Bluetooth uses a distributed pull-based approach for service advertising and discovery. To this end, each device maintains its own service discovery database (SDDB), which is a registry where its services are advertised. Therefore, a Bluetooth device performs service discovery by querying the SDDBs of the devices around. These advertising and discovery processes are illustrated in Figure 5. Notice that, initially, all devices advertise their services on their respective SDDBs (1). Next, a client searches for all the Bluetooth devices on the range of its wireless interface (2). For each device found, the client sends a query about the availability of an interested service (3). The devices answer these queries by informing whether they offer the needed service or not (4). Once localized a device providing the desired service, the client can connect directly to such device and finally use the service (5).

**FUTURE RESEARCH TRENDS**

Although the first service-oriented pervasive computing solutions have been developed, much work has to be done yet. For example, the matching of the user’s needs and the services’ functionalities should be enhanced to improve the interaction of the user with the pervasive application. Still, problems remain in the context of service continuity. Solutions to this problem would enable the user to use a service continuously, as he/she walks through different environments. This is important because, sometimes, the service a client was using is not available in the new environment, and thus, some mechanism should allow it to use a similar one without, or at least with a minimum, of interruption.

These, and possibly other problems related to the provision of services in pervasive environments, must certainly be completely solved so that we can enjoy the full potential of merging service-oriented and pervasive computing.
CONCLUSION

The service-oriented paradigm has proved to be an important element in pervasive computing systems, in order to provide anytime and anywhere access to services. Its dynamic binding feature enables to build applications powered with on-demand extensibility and adaptability, two important elements of any pervasive system.

Given this trend, in this chapter we have tried to present an overview of service provision in pervasive computing environments. More precisely, we have showed an introduction to the main characteristics, challenges, and solutions concerning the way that services are advertised, discovered, and used in pervasive environments. Although we presented concepts at an introductory level, we believe they may serve as a good source of knowledge, helping both students and researchers involved with these fields.

REFERENCES


Service Provision for Pervasive Computing Environments


KEY TERMS

Pervasive Computing: The vision conceived by Mark Weiser which consists of world where computing will be embedded in every day objects.

Service: A software entity that can be integrated to a remote distributed application.

Service-Oriented Computing: The newest paradigm for distributed computing, where applications should be built by dynamically integrating services.

Service Advertisement: The element used for publishing and discovering a service.

Service Client: The one wishing to use a service.

Service Provider: The one that offers services.

Service Registry: The place where services are published.

This work was previously published in the Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 877-884, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 1.22
Preparing Participants for Computer Mediated Communication

Robert Jordan
U.S. Bureau of Labor Statistics, USA

ABSTRACT

Computer mediated communication (CMC) provides a way of incorporating participant interaction into online environments. Use of such features as discussion forums and chats enhance collaborative work and learning. For many, however, CMC may be an unfamiliar medium. To ensure a successful CMC event, it is essential to adequately prepare participants for CMC. A proposed four step model prepares participants for CMC. The four steps include conducting a needs and population analysis, providing an orientation before the event and shortly after the event begins, and providing continuing support.

INTRODUCTION

Computer mediated communication (CMC) provides interaction during online events, whether synchronous or asynchronous. The moderator-to-participant and participant-to-participant interaction present in CMC offers participants an environment that closely replicates face-to-face interaction with which they may be more comfortable. The advantages of CMC are diminished if participants are not adequately prepared for its deployment and use. In addition to examining why participants may be uncomfortable with CMC-based interaction, this paper proposes several strategies for effectively preparing participants for CMC. These strategies are based on experiences delivering online training across the country to an audience that was unfamiliar with CMC-based interaction.

BACKGROUND

Moore and Kearsley (2005) cite one common misconception that participants new to online learning hold is that learning online is less demanding than face-to-face learning. In fact, the
Preparing Participants for Computer Mediated Communication

opposite is generally true. Effective participation in CMC environments requires participants to be highly motivated and less passive than what may be possible in a face-to-face environment. They must be prepared for what they will experience in CMC. If not addressed at the outset, this misconception may result in participants falling behind and never becoming comfortable and engaged in CMC.

Sherry (2000) cites Fishman in highlighting several potential constraints on CMC-based learning. These constraints include the participant’s lack of experience with using computers, cultural influences, and apprehension about electronic communication (Sherry, 2000). Hiltz and Wellman (1997) describe a sense of pressure on new CMC participants and an uncomfortable feeling of being lost that many of them experience after they first venture into a CMC environment. In addition, new participants are often apprehensive about using technology to communicate and may be more guarded in what they post in discussion forums or chat rooms when compared to what they might otherwise communicate in a face-to-face environment (Puetz, 2000; Sherry, 2000). They may need time to find their online voice (Sherry, 2000). Finding one’s own voice, knowing what to say and when to interject it in an online environment is critical in CMC and the earlier that participants are comfortable with this, the more successful the CMC event will likely be. As features such as instant text messaging and chat are more widely incorporated in everyday life, participants should be more comfortable with CMC.

Many universities incorporate instant messaging in their online education. One such example was the use of MSN Messenger in a mathematics course at the University of Southern Queensland (Loch & McDonald, 2007). Students in the mathematics course were comfortable using many of the features of the instant messaging application, including the electronic ink feature (Loch & McDonald, 2007). Increased organizational use of instant messaging and its improving functionality will likely result in its increased use in CMC events.

While a lack of comfort in communicating online is the primary challenge for CMC, surveys have indicated several other factors that may result in discomfort with CMC (Lewis & Romiszowski, 1996). Some participants are uncomfortable with the nature of asynchronous communication and may be frustrated with the time lags between postings and responses in online discussion forums (Lewis & Romiszowski, 1996; Sherry, 2000). For some participants, the delay in back-and-forth communication lacks the spontaneity and directness of face-to-face communication. Some participants are reluctant to express candid personal views in discussion forums or chats, since they fear that their postings can easily be printed out or saved (Lewis & Romiszowski, 1996). Apprehension about leaving a “paper trail” is understandable. What might be said and forgotten in a face-to-face environment may be misinterpreted and taken out of context in a written form. Participants may become frustrated when others post comments in discussion forums and in chat sessions that reflect their own ideas before they get a chance to comment (Lewis & Romiszowski, 1996). In a face-to-face environment, participants may be able to better gauge when they should interject comments. Those who post late in discussion environments are at a disadvantage in expressing what might appear to others to be original ideas. This situation occurs in face-to-face environments as well, as anyone who is familiar with the overeager “hand raiser” can attest. It might seem to be more clearly evident in CMC due to the text-based method of communicating. The give-and-take of face-to-face verbal interplay is replaced with a written record of who stated what and when.

Technology continues to evolve and may mitigate some of these challenges. For instance, some
discussion board applications now allow users to easily post audio or even video postings and responses, thus reducing the emphasis on text-based communication. In addition, most CMC tools allow participants to “subscribe” to discussion board forums so they can receive instant updates of activity on the boards either via e-mail or really simple syndication (RSS) site feeds.

Another concern expressed about CMC is that online discussion board or chat postings, made without the benefit of non-verbal visual cues, may be misinterpreted and more contentious than what might otherwise occur in a face-to-face environment. There are social communication mores that are evolving in online environments, such as the wider usage of emoticons, which are alleviating this concern. Many available Web conferencing and other CMC tools now formally incorporate the use of emoticons as non-verbal cues in communication.

Moore and Kearsley (2005) cite four reasons for participant anxiety in online environments.

1. Participants are apprehensive about CMC, particularly those new to it
2. Some participants express concern about the quality of online events, such as courses
3. Access to technology-based resources necessary for participating may be a challenge for some participants
4. Some participants worry about the personal resources and expenditures that are required to fully participate in CMC

Technological challenges remain. Participants who can not easily access CMC applications are at an immediate disadvantage from which they may never fully recover. For example, some tools require downloadable plug-ins and players, which may pose security problems if the participant is using a computer supplied by their workplace or public library. Bandwidth availability must be considered as well if the tools used are rich in multimedia. As use of Web-based applications expands, these challenges should diminish. For the present, however, developers of CMC events are wise to assess these potential challenges.

STRATEGIES FOR INCREASING PARTICIPANT COMFORT WITH CMC

Lorraine Sherry (2000) suggests five keys to adoption of CMC by participants:

1. They must believe there is a clear benefit to engaging in CMC
2. They must overcome any fear they have of technology
3. They must feel social and cultural compatibility in the CMC environment
4. They must feel comfortable with the CMC event’s “scaffolding”
5. They must find their online voice

Sherry (2000) cites Engestrom in highlighting the importance of cultural effects. Ultimately, CMC is a social culture that includes those who participate in it, the online tools and conventions that foster it, and the tangible and intangible results that accrue from it (Sherry, 2000). For CMC to be successful, participants must believe that they have a stake in the process and feel a connection with fellow participants as could exist in a face-to-face interaction. As online interaction expands, sustaining online communities should become easier, but resistance remains strong among those who view online communities as remote or impersonal.

There are several strategies for preparing participants for online participation. Sherry (2000) cites Wilson and Ryder who suggest several strategies. These include demonstrating the technology prior to a CMC event, providing online consultation with participants, employ-
Preparing Participants for Computer Mediated Communication

Preparation for moderated discussion forums, establishing a model for “netiquette,” and providing desktop aids and tip sheets for participants (Chou, 2001; Sherry, 2000). Such tactics reassure CMC participants and reinforce the idea that they are part of a larger community.

Gilly Salmon (2003) recommends an online orientation session highlighting tips for communicating before online courses using CMC. Orientation might take the form of a face-to-face demonstration of how CMC works or in a Web conference, in which the system and tools to be used in the CMC event can be demonstrated and “remote control” passed to participants so they can gain comfort in using the tools. In addition, moderators might decide to incorporate features such as online seminars, virtual office hours, and participant journals as ways to make participants more comfortable with CMC (Chou, 2001). Including chat sessions creates an interactive environment that more closely resembles face-to-face interaction. Emerging Web conferencing tools provide a convenient and effective way to build participant confidence in CMC technology before they enter a CMC event. Many of the Web conferencing tools currently available provide polling tools that allow moderators to assess the feelings of participants toward CMC before the event occurs and allow the moderator to make adjustments if needed. Web conferencing applications, such as WebEx, Adobe Connect, Live Meeting, or Centra among others, are generally more effective than the chat tools currently often integrated in asynchronous online environments.

Developers of CMC-based events should integrate a variety of learning styles into their events by providing for a variety of participant activities (Salmon, 2003). The advantage to incorporating various learning styles is that it promotes the concept that there is something to accommodate all participants in the CMC environment. An auditory learner, for example, might appreciate an audio or video clip introducing the event. A visual learner might appreciate well-designed tables or simulations and screen captures with voice-over narration. Tactile learners might find online collaborative exercises and projects engaging. In any CMC event, moderators should strive to keep the entire audience interested in not only the communication but also in the content on the site. An interested participant is more inclined to contribute.

CMC moderators should engage new participants quickly to make them comfortable with communicating online. This particularly is true for inexperienced participants, especially those who may be skeptical or fearful about participating in CMC. The optimal solution is that participants are comfortable with logging into an event. Preparation should begin well before the event and continue as the event unfolds. The next section will describe a specific plan for preparing participants for CMC, with an emphasis on preparing the most inexperienced participants and increasing their comfort with what will be expected of them in a CMC environment.

A MODEL FOR PREPARING PARTICIPANTS FOR CMC

A four-step model assists in preparing participants for CMC events. The four steps are as follows:

1. Analysis of needs and population
2. Participant CMC orientation (before event)
3. Online introduction to CMC
4. Continuing support

Analysis of Needs and Population

As in many development processes, the analysis phase is essential. Developers must analyze the needs of the specific population that will be served to determine the appropriate levels of the
Preparing Participants for Computer Mediated Communication

orientation and continuing support necessary to ensure success of the CMC event. Performing a thorough analysis in advance of the event will provide information to the developers that they can use in building their event. The model for providing analysis discussed here is based on two models that are often incorporated in traditional instructional system design models:

1. Needs assessment
2. Population analysis

Needs assessment reveals the attitudes of the targeted audience as well as their needs (Hodell, 2000). For online training using CMC, this step would likely be undertaken as a part of the instructional systems design process, but is should be completed for any significant CMC event. While the focus of the needs analysis will often be focused more on the subject matter content that will be discussed during the event, the information is of value in assessing how CMC can be effectively used to deliver the subject matter. For instance, if the needs assessment reveals a desire for participant and moderator interaction revolving around specific topic areas, the event can be designed with discussion boards or synchronous chat sessions. What developers should be looking to obtain from the needs analysis is how CMC will be employed during the event to meet the interests of the audience and build upon their needs.

Population analysis assesses the targeted audience (Hodell, 2000; Puetz, 2000). This is important to the CMC developer and moderator in that it will determine the direction of the remaining steps in the four-step model highlighted above. Questions in the population analysis survey should be designed to assess the comfort level of the targeted audience with the methods commonly used in CMC (Puetz, 2000). Hodell (2000) recommends building a population profile of the targeted audience from the results of the survey. Surveys might include the following questions:

1. How much time is spent online during an average work week?
2. How much time is spent online outside of work during an average week?
3. Are CMC features such as instant messaging, chat rooms, and discussion boards commonly used in the audience’s workplace?
4. How familiar and comfortable is the targeted audience with these features and do they use them frequently at home and in the workplace?
5. What is the audience’s perception of online environments versus face-to-face environments?

Taken together, the needs assessment and population analysis provide an important first step in the process of preparing participants for CMC. Without completing the analysis step, a CMC event developer will be at a disadvantage.

Participant CMC Orientation (Before Event)

The level of participant orientation necessary before a CMC event should be determined once the results of the assessment in step one are complete. Orientation should begin one to two weeks before the event. The objective of the orientation is to provide participants with a clear idea of what to expect when they first logon to the event and familiarize them with the features that they will be expected to use to communicate with fellow participants.

Two orientation methods are suggested:

1. Asynchronous and synchronous online orientation
2. Desk aids

Online orientation may be either asynchronous or synchronous. Asynchronous orientation provides participants with a virtual tour of the features of the technology that will be employed
Preparing Participants for Computer Mediated Communication

during the event. The CMC technology that will be used in the event may also have a tutorial that can be provided to participants. There are many screen capture applications available in which screen actions are recorded with or without voice-over narration. Recorded actions should include a demonstration of the features that will be used during the event.

Synchronous orientations sessions provide the additional benefit of participant interaction. Commercial Web conferencing tools allow moderators to demonstrate the features of the system being used to participants. Web conferencing also allows for participant questions and reactions which may be useful in refining the event before it occurs.

Desk aids that participants can refer to throughout the course are also invaluable. Participants can refer to the desk aid until they become familiar with the features of the CMC system that will be used in the event. The desk aid might incorporate screen captures that show participants how the screens look when they log into the CMC event. Other functions such as logging in, posting in discussion forums, and participating in chat sessions can be included in the desk aid. Policies and restrictions may be included in the desk aid as well. The desk aid may be distributed in electronic format so participants can print it out as needed.

**Online Introduction to CMC**

Participant preparation models exist, including a five-step model for online learning developed by Gilly Salmon (2003). Salmon’s model is applied as the CMC event begins. The first two steps of Salmon’s five-step model directly relate to initial orientation. These two steps are access and motivation and online socialization (Salmon, 2003). In the first week, moderators must ensure that participants access the system successfully. If there are challenges initially in accessing the event, participants may drop out of the event or never become fully engaged with the event. This is particularly true for those participants who are not comfortable with technology or the demands that are made of CMC participants. To ensure full participant access, moderators should monitor participation and contact participants who do not log in shortly after the event begins. Moderators should also provide an e-mail address or a telephone number so participants can contact them if they experience difficulties accessing the event.

Online socialization early in the event is desirable. It is suggested that a discussion forum be established for participant introductions. The moderator should encourage participants to post early during the event. Participants should provide information on themselves, their reasons for participating in the event, and perhaps post a photograph. Building online socialization early in the event will promote communication throughout the event.

Another suggestion for early in the event is to provide participants with information on proper techniques for communicating in an online environment. Articles on “netiquette” may be attached or linked to from within the CMC Web site for participant access.

**Continuing Support**

Participants should be supported throughout the course of the event. It is suggested that a discussion forum be established for questions that the participants might have. Moderators should be vigilant in monitoring participation and be prepared to intervene if participants fail to contribute. It is possible that a lack of participation may indicate that participants are experiencing technical or other difficulties.

**CASE STUDY**

The techniques described above were used in delivering an online training based on CMC.
that was delivered over a four week period. The application used for this training was an open source courseware management system called the modular object oriented developmental learning environment, or Moodle. Moodle provides functions such as discussion forums and chat rooms commonly used in CMC. Well before the development team selected Moodle as the tool to deliver the training, work had begun to prepare participants for CMC.

**Analysis of Needs and Population**

Analysis of the population of potential learners and their needs began before the training was set to begin. The population was comprised of approximately 170 participants from nearly every U.S. state, the District of Columbia, the Virgin Islands, and Puerto Rico. A survey was developed to assess the readiness of this population for CMC and their needs, with a focus on what features of CMC they wanted to see included in the online training.

The survey showed the training developers that the majority of survey respondents frequently used the Internet in their work and were comfortable with computer-based technology. In addition, a number of the respondents had participated in online learning events with CMC features. The respondents of the survey indicated a definite interest in interaction with online moderators as well as with fellow learners.

The conclusion from this front-end analysis was that the population was comfortable with using the Internet and had some experience with online learning and CMC, but by no means had mastered the functions of CMC. The audience did express interest in CMC features that would enable interaction with the course moderators and fellow participants. These results assisted the online training developers in approaching the orientation.

**Participant CMC Orientation (Before Event)**

A week before the event began, participants were provided with a training plan and an orientation desk aid. The desk aid covered navigating to the training Web site, establishing a Moodle account, and signing into a specific section of the training. Also included was information on participating in discussion forums and chat sessions. Screen captures from the Moodle site with detailed explanations were included in the desk aid. Web conferencing might have been used as well, if one had been available to the course developers.

**Online Introduction to CMC**

In the first week’s activity, participants were provided a discussion forum entitled “Ask the Moderators,” where they were encouraged to ask questions related to the course. Moodle incorporates an internal messaging system in which participants sent direct messages to the event moderators. The moderators’ work e-mail addresses were also provided. Finally, a link to a Wikipedia entry on “netiquette” was posted in the first week’s activities.

As a result of the preparation the great majority of participants experienced no difficulty logging into the Moodle Web site. In the few instances where participants experienced difficulties logging in, the difficulties were resolved quickly, in part because they knew how to quickly contact the moderators. Participants indicated in the evaluation sheets that the desk aid helped them to log in and know what to do once they were in the Web site. They also appreciated knowing the moderators could be contacted easily from within the Web site.

**Continuing Support**

The “Ask the Moderators” discussion forum was included in all four weeks of activities. The mod-
Preparing Participants for Computer Mediated Communication

ERATORS made use of a Moodle feature referred to as a “new forum” discussion board to post reminders and tips for enhancing the experience.

FUTURE TRENDS

As more organizations look to contain costs, it is expected that even more of them will turn to some form of CMC. As CMC draws more participants, there is potential for problems if these new users are not fully prepared for what is required in a CMC environment. With more people using Internet technology, including CMC, it can be expected that less intensive preparation may be required in the future. The nature of the preparation may also change with the evolution of technology and in the way CMC is delivered. For instance, audio-based discussion boards now exist allowing participants to record their postings and responses to posts, rather than responding by typing text.

Another future trend is the emergence of more sophisticated modes of online communication. One popular social online community is called Second Life, a virtual reality gaming environment that allows participants to create alternate identities and lifestyles through avatars. Community members can buy and sell virtual products, services, and even real estate using virtual dollars that can be exchanged for real money (Newitz, 2006). The success of these virtual communities serves as a model for creating and sustaining an online community. Companies such as Wal-Mart, American Express, and Intel have been studying interaction in Second Life in conjunction with their corporate training objectives (Newitz, 2006). Ohio University has been incorporating Second Life into its formal online learning programs (Dewitt, 2007). Virtual communities such as Second Life provide a more life-like form of interaction than what is generally seen today in CMC.

The closer developers can get to replicating face-to-face interaction, the more likely that CMC will be accepted by a broader audience. Virtual reality applications, such as those provided in Second Life, provide a glimpse into the future of CMC can become.

CONCLUSION

Organizations are turning to CMC as a way of meeting, collaborating, and learning. More people are being introduced to CMC, including users who are unfamiliar with the style of communication or the technology involved. Participant acceptance of CMC will grow if they are prepared for what is encountered in a CMC environment.

A four-step process may be employed to prepare participants for CMC. These steps include conducting a needs and population analysis on the target audience, providing a participant orientation before the CMC event begins and during the first week, and providing continuing support throughout the event. A systematic process ensures that CMC developers and moderators will understand their audience’s strengths and weaknesses and will enable them to develop an effective deployment strategy. Communication techniques found in CMC offer participants an environment that more closely replicates face-to-face interaction with which they may be most comfortable. Preparing participants for CMC is critical in making participants comfortable with communicating online.

REFERENCES


Preparing Participants for Computer Mediated Communication


KEY TERMS

Asynchronous: Online communication that does occur at the same time.

Avatar: A computer-generated representation of an online participant commonly found in virtual reality environments.

Computer Mediated Communication (CMC): Design that incorporates synchronous and asynchronous online communication that promotes interaction among participants. Such features include discussion forums and online chat features.

Needs Assessment: An analysis of the needs and desires of a target audience. Methods of conducting a needs assessment include survey questionnaires, focus groups, and personal interviews.

Open Source: Programs that are developed with an open license and distributed without cost. Upgrades to programs and support are often provided by the user community.

Population Analysis: An analysis of a targeted audience focusing on its attributes, abilities, and feelings.

Synchronous: Online communication that occurs among participants simultaneously.

This work was previously published in the Handbook of Research on Computer Mediated Communication, edited by S. Kelsey and K. St.Amant, pp. 25-33, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 1.23

Computer–Mediated Communication Research

J. D. Wallace
Lubbock Christian University, USA

ABSTRACT

This chapter asks “What is meant by computer-mediated communication research?” Numerous databases were examined concerning business, education, psychology, sociology, and social sciences from 1966 through 2005. A survey of the literature produced close to two thousand scholarly journal articles, and bibliometric techniques were used to establish core areas. Specifically, journals, authors, and concepts were identified. Then, more prevalent features within the dataset were targeted, and a fine-grained analysis was conducted on research-affiliated terms and concepts clustering around those terms. What was found was an area of scholarly communication, heavily popularized in education-related journals. Likewise, topics under investigation tended to be education and Internet affiliated. The distribution of first authors was overwhelming populated by one time authorship. The most prominent research methodology emerging was case studies. Other specific research methodologies tended to be textually related, such as content and discourse analysis. This study was significant for two reasons. First, it documented CMC’s literature historical emergence through a longitudinal analysis. Second, it identified descriptive boundaries concerning authors, journals, and concepts that were prevalent in the literature.

INTRODUCTION

Computer-mediated communication (CMC) involves a wide number of characteristics involving human communication. It also includes systems, methods, and techniques that are typical of online environments. Therefore, one would rightfully expect definitional difficulties both technological and methodological. Wallace (1999) extensively surveyed the literature concerning CMC and found relatively few definitions. While differences abounded in the definitions found, the one constant was the use of the computer as an intermediary device. The centrality of the computer and communication layers human characteristics and technological issues.
Levinson (1990) suggests that in order to understand a device or a technique, not only should we take a microscopic view through research and examination, but we should also take a more macroscopic view. A survey of the scholarly communication might help provide a different perspective. Hopefully, it would reveal some of the larger areas of inquiry concerning online research in general, and computer-mediated communication specifically. This macroscopic survey would enable researchers and scholars to more efficiently coordinate their own activities with outlets and concepts that have the most pressing need for their contributions. It has the additional benefit of documenting CMC's developmental features that can be compared with future research or differing methodologies.

Similar to other such studies, the purpose of this chapter is provide an overview of the CMC scholarly literature, and to “identify its component features in providing a tangible means of identification” (Dick & Blazek, 1995, p. 291). Likewise, it is not to determine the magnitude that CMC occupies as a discipline, field, specialty, or subspecialty area. For purposes of literary description, the term “field” is not a cataloguing designate, but rather a convenient moniker under which CMC scholarship resides. CMC is often described in the literature as a field. However, designates of specialty, or subfield are probably more accurate.

Simply put, the statement of the problem is: “what are trends in computer-mediated communication research?” Definitions and descriptions of current literature on the subject reflect views that are selective and often disparate. Rather than revisit debatable definitional issues, an arguably more objective approach will be used as the focus of this inquiry. Specifically, what authors, journals, concepts, and research issues possibly populate the CMC domain?

Certainly, a number of conceptual problems would be introduced with any kind of predictive examination (Hargittai, 2004). Therefore, exploratory and descriptive procedures seem more appropriate than postulating hypotheses. With this in mind, the original question concerning CMC has, as one possible answer, a bibliometric analysis into the nature of the field. Bibliometrics is the “… mathematical and statistical analysis of patterns that arise in the publication and use of documents” (Diodato, 1994, p. ix).

Library and information science have long used bibliometrics for this kind of analysis. They have a body of literature supporting their validity and reliability. Moreover, bibliometric procedures provide the means upon which relationships of theoretical inquiry can be based. Borgman & Rice (1992) state that: “Bibliometric data are particularly useful for studying longitudinal trends in scholarly disciplines because of the massive datasets that can be utilized. Virtually no other method provides as comprehensive coverage of a topic in scholarly communication” (p. 400).

Journal articles appear to be reasonable and available artifacts for identifying this area. This is done for three reasons. First, their affiliation with the bibliometric theory of problematic network analysis provides support for their representation of scholarly activity. (Coutial, 1994; Courtial, Callon, & Sigogneau, 1984). This theory views scholarly communication and literature as a series of problematisations that reflected the underlying intellectual discourse. Journal articles are seen as the primary artifact from which to extract elements of that discourse. Second, analyses of journals do not consume the financial resources inherent in more exhaustive treatments. Third, their online availability makes this current analysis easier to replicate and contrast when used as a benchmark upon subsequent research.

Research questions. One advantage of surveying the field is to let the “field” define itself as it exists in the extant literature. In this regard, almost 40 years of archival data was examined by combining the results of a previous survey (Wallace, 1999) with this 2005 examination. Hopefully, this will offer a more comprehensive look at how CMC
research is viewed through a database analytic lens. As search engines replace annotated bibliographies and other indexical instruments, the nature of the scholarly profile produced through this lens becomes more critical.

In regard to the above literature framework, this chapter poses four research questions concerning the computer-mediated communication research. These questions are general in nature, and align themselves theoretically with artifacts, producers, and research concepts. The conceptual issues are somewhat more complex than general identification and demarcation that led to the creation of a fourth research question. Simply put, the research questions posed are:

- **RQ1:** What are the primary journals of CMC literature, both past and present?
- **RQ2:** Who are the producers of CMC literature, both past and present?
- **RQ3:** What are the CMC concepts being examined in scholarly journals?
- **RQ4:** How do the current research concepts of CMC relate to other topics in the field?

**Research question 1: Journals.** Research question 1 addresses the scholarly journals representing literature concerning computer-mediated communication. Specifically, it will identify these artifacts by the analysis of frequency data and an identification of core journals across several disciplines. Common databases were set up both in the previous and the current study so that frequency and other component features could be identified.

One advantage of using journals over other artifacts, such as citations, is that they are believed to contribute to minimizing threats to validity (Cronin & Overfelt, 1994; Glaenzel & Moed, 2002; Hyland; 2003; Schloegl & Stock, 2004). Citations often go to secondary sources (obfuscating the real influence), and authors often do not hold the opinions that citation patterns attributed to them. Once frequency counts are established, a number of other component features can be derived, such as prolific and core journals. This may provide indicators of publication patterns that can be used to target future scholarship.

**Research question 2: Authors.** Research question 2 looks at producers of CMC’s scholarly communication. It will be addressed by the analysis of frequency data concerning authorship. Frequency of author publication rate will give a glimpse into the most prolific authors, and has been indicative of some of the major theoretical contributions of bibliometrics (Cronin & Overfelt, 1994; Nicholls, 1989; Rice, Chapin, Pressman, Park, & Funkhouser, 1996).

**Research question 3: Concepts.** Research question 3 identifies the more prominent areas of focus in regard to CMC research. It will be addressed through the analysis of indexical key words from the master database. Indexical key words were chosen as concepts because they have demonstrated useful in identifying research trends in other studies (Callon, Courtial, & Laville, 1991; Cambrosio, Limoges, Courtial, & Laville, 1993; Stegmann & Grohmann, 2005). Key words stand for indicators of scholarly problematisations that are being articulated in the literature (Courtial, 1994; Callon et al., 1991)

**Research question 4: Relationship of research concepts.** Research question 4 investigates how the research concepts of CMC relate to topics in the field. It will be addressed through co-word analysis. This is generally done by analyzing the proximity that key words and other concepts have with one another (Diodato, 1994; Larsen & Levine, 2005; Stegmann & Grohmann, 2005).

**METHOD**

In order to produce a literature survey that spans almost 40 years, two studies were combined. A more recent analysis of CMC research extended a previous CMC literature study (Wallace, 1999). The previous study’s data was collected in June
of 1998, for 1997 and prior, utilizing the indexes of ABI/INFORM, ERIC, PsycInfo, Social Science Index, and Sociofile. The more recent study’s data was collected in August of 2005. EBSCO was used as a common vender in the 2005 study. Academic Search Complete, Communication & Mass Media, Business Search Premier, ERIC, PsychInfo, and PsychArticles were selected as databases. The selected databases are considered prominent resources in the areas of business, education, psychology, and the social sciences. Furthermore, these databases are noted for their strong research affiliations, and have an intuitive link with CMC. A key word search of the respective databases was done for the variable “computer-mediated communication..” Both hyphenated and unhyphenated variations of the term “computer-mediated” were used to insure breadth of coverage.

Each of the databases had idiosyncratic characteristics that made identical survey techniques, across all databases, impossible. However, all methodologies adhered to the following guidelines. Databases were examined to see if there were any inherent journal markers that would enable the extraction of journal articles. Books, book reviews, conference papers, and other nonjournal materials were eliminated through either filters or inspection. Extraneous information and fields were eliminated. Documents were then placed into a common file. Procedures were similar for both studies, and a relatively exhaustive description is available in Wallace (1999).

**Data Analysis**

**Journals.** The data file was examined in terms of journal frequency. From this, it could be determined the identity of CMC journals and how they were positioned in terms of the overall literature as defined by this chapter. Subsequently, Bradford-type partitions were derived to identify the core journals.

Bradford partitions are where journals are ranked from most to least prolific in terms of number of articles produced concerning the subject under scrutiny. They are then divided into “zones” that have roughly the same number of articles. Predictably, the journals that produce the most articles have the least number of journals in their zone. The zone that has the least journals is used to identify the “core” journals for a subject area. Bradford recommended the use of three zones (Diodato, 1994). The previous study serendipitously arrived at three zones through a somewhat more complex procedure. These zones had a relatively equal number of journal articles with the core zone containing a relatively few prolific journals. For comparative purposes, the more current study utilized three zones as well.

**Authors.** The 2005 analysis of authors was also restricted by comparative constraints. The previous study used the straight count method of authorship. The straight count method identifies all the first authors from the journal database. Therefore, both studies used straight count to extract authors for analysis.

After authors were extracted from the master database, they were then rank ordered, and standardized. Frequency and cumulative data was then calculated. The CMC authors were then analyzed from an “authors x number of contributions” format.

**Concepts.** Frequency of problematisations overall were calculated and analyzed for segmentation. This is not only critical for this research question, it is imperative for RQ4. Both Hinze’s (1994) “10 journals or more” and Rice et al.’s (1996) “top 200 key words” were considered exemplars in this regard. The decision was made to view the distribution in terms of which kind of demarcation was most efficient while capturing the majority of the distribution. Fifty percent was considered the benchmark on which to base the comparisons. This is an area where the literature distribution must be known to be able choose the more efficient
method. If neither of the previously mentioned methods surpassed the 50% benchmark, then the literature would be deemed too disparate to determine more prominent terms.

Once segmentation was determined, then the core key words were longitudinally examined for their visibility. The previous study examined key words concerning 1997 and prior. The more recent study surveyed the literature from 1997 through 2005. The overlap was intentional as many databases' literatures are not well represented in their last year of analysis (Egghe, 1990). Indicators of this “droop” were present in the previous 1997 survey.

Relationship of research concepts. The above concepts were then scanned for more familiar research affiliated terminology. Terms that appeared methodological, or had the term “research” as a component-attached term were extracted. Establishing relationships was done through cluster analysis. A number of clustering procedures have been used in the analyses of emerging fields. The complete linkage method (furthest neighbor) was considered an appropriate procedure because of its rigor over the single linkage method and its comparability to the previous study (Aldenderfer & Blashfield, 1984; Hinze, 1994, Wallace, 1999). The lack of knowledge of current granularity of findings also makes this procedure well suited. (McCain, 1989; Spasser, 1997). As with other studies, similarity between concepts was measured using a Pearson relation procedure. (e.g., Hinze, 1994; McCain, 1989; Spasser, 1997).

RESULTS

The 1997 CMC study generated 611 unique article references for 1997 and prior. Four hundred and fifty-nine, or about 75%, tended to be education related as indicated by their ERIC affiliation. The current examination started from a more mature 1997 dataset through 2005. It generated 1,326 unique article references. Nine hundred and fifty-three, or about 72%, tended to be education related as indicated by their ERIC affiliation.

The following sections survey journal publication trends for CMC, identify some of the more common terminology within that survey, and then narrow that terminology to research-affiliated terms and concepts clustering around those terms.

Artifacts: Survey of Core Journals

A Bradfordian-type analysis was done both in the more current and the previous study (Table 1). In the 1997 study, journals had to produce six or more articles total to be considered core. In the 2005 study, journals had to produce 11 or more articles.

While roughly the same number of journals were considered core in both studies, only six journals were recognized as core in both studies. These were American Journal of Distance Education, Computers & Education, Computers and Composition, Educational Media International, Information Society, Internet Research, Journal of Educational Computing Research, and TechTrends. A cursory look at journal affiliation reveals that the journals are largely education affiliated. Two notable exceptions are Information Society and Internet Research (see Table 2).

<table>
<thead>
<tr>
<th>Zones</th>
<th>Early Journals</th>
<th>Early Articles</th>
<th>Recent Journals</th>
<th>Recent Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22*</td>
<td>195</td>
<td>25*</td>
<td>430</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>196</td>
<td>101</td>
<td>427</td>
</tr>
<tr>
<td>3</td>
<td>153</td>
<td>195</td>
<td>360</td>
<td>426</td>
</tr>
</tbody>
</table>

*6 or more articles

*11 or more articles
Productivity of these core journals ranged, in the earlier research, from a high of 16 to a low of 6. The more recent research frequency was appreciably higher ranging from a high of 34 to a minimum threshold level of 11.

To give an idea how this distribution might be comparable to other literature descriptions, in terms of longevity, it was tracked over time. While no journals appeared till 1984, databases were searched in previous years. The most far-ranging search occurred in ERIC, which started in 1966. Records were also examined from ABI/INFORM, Psychlit, Sociofile, and Social Science Index in 1986, 1971, 1971, and 1983, respectively. All databases were producing articles by 1986.

Producers: Survey of Core Authors

Three or more references were used to demarcate core authors in both studies. This formulation has the benefit of fitting within the general contention that authors producing one or two articles have the vast majority of the distribution and should be excluded from consideration (e.g., 94% by Burnham, Shearer, & Wall 1992; 83% by Keenan, 1988). On the other hand, because the production numbers were low in general, using more than three was considered too severe a threshold for the respective studies.

The CMC literature examined was heavily weighted toward authors producing only one article. The early 1997 research had 506 different first authors producing 611 different references. Close to 96% of the authors had one or two publications. They accounted for 87% of the articles produced. The 21 core authors were published in 54 different journals producing 80 different articles. The 2005 survey had a similar distribution, with 1,169 authors producing 1,326 different articles. However, the core authors produced a somewhat smaller 8.5% of the articles. Specifically, there were 31 authors that produced 113 articles in

### Table 2. Survey of core journals for CMC

<table>
<thead>
<tr>
<th>1966-1997</th>
<th>1997-2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behaviour and Information Technology</td>
<td>British Journal of Educational Technology</td>
</tr>
<tr>
<td>Canadian Journal of Educational Communication</td>
<td>Business Communication Quarterly</td>
</tr>
<tr>
<td>Communication Education</td>
<td>CALICO Journal</td>
</tr>
<tr>
<td>Communication Research</td>
<td>CyberPsychology &amp; Behavior</td>
</tr>
<tr>
<td>Computers in Human Behavior</td>
<td>Distance Education</td>
</tr>
<tr>
<td>Educational Technology</td>
<td>Distance Education Report</td>
</tr>
<tr>
<td>Educom Review</td>
<td>Indian Journal of Open Learning</td>
</tr>
<tr>
<td>Human Communication Research</td>
<td>Instructional Science</td>
</tr>
<tr>
<td>Interpersonal Computing and Technology</td>
<td>Internet and Higher Education</td>
</tr>
<tr>
<td>Journal of Communication</td>
<td>Journal of Adolescent &amp; Adult Literacy</td>
</tr>
<tr>
<td>Learning and Leading with Technology</td>
<td>Journal of Educational Technology Systems</td>
</tr>
<tr>
<td>Organization Science</td>
<td>Journal of Instruction Delivery Systems</td>
</tr>
<tr>
<td></td>
<td>Journal of the American Society for Information Science</td>
</tr>
<tr>
<td></td>
<td>New Media &amp; Society</td>
</tr>
<tr>
<td></td>
<td>Quarterly Review of Distance Education</td>
</tr>
<tr>
<td></td>
<td>Small Group Research</td>
</tr>
</tbody>
</table>
71 different publications. Only Joe Walther was included in both previous and recent sets of core authors (Table 3).

**Concepts: Survey of Key Terms**

First, to contextualize the conceptual survey concerning CMC, it is useful to understand the scope of the current and previous examination. The previous study had 1,787 unique problematisations (indexical terms) contributing to a total dataset of 6,898 terms. The 2005 study had 2,700 unique problematisations contributing to 12,935 total terms.

The overall datasets were then examined for more dominant concepts. Both Hinze’s and Rice’s methods detailed earlier surpassed the 50% threshold. The previous study used Hinze’s (1994) demarcation, and identified 116 core terms. They accounted for more than 3,684, or 53%, of the total indexical terms mentioned. However, Hinze’s demarcation was primarily used because the previous study’s emphasis was on the totality of the conceptual base. Since the 2005 study was designed specifically to identify research affiliations within that base, Rice et al.’s (1996) more liberal “top 200 terms” was utilized. This accounted for 7,827 or roughly 60% of the entire conceptual set.

While a complete discussion of the core terminology would be unwieldy at best, a physical inspection of the two studies indicated about 30% of the totality of terms is included in the 21 most prominent descriptors. With the admitted concession that computer-mediated communication is the search term predicating the study, these terms still provide an overview of some of the more dominate themes that have emerged in the CMC literature. Table 4 details both studies with the count and rank of terminology that occurred.

Thirteen of the 21 terms appeared in both studies. As would be expected, even with rank differentiation, the more recent study has substantially more contributions. Online systems had the highest overall percentage increase. Decreases were only observed concerning the terms electronic mail and computer networks.

There were eight terms that were not reciprocal (Table 5). The overwhelming number of these fell in the bottom half of the rankings where predominance was not as stable.

---

**Table 3. Survey of CMC first authors**

<table>
<thead>
<tr>
<th>1966-1997 Authors</th>
<th>1997-2005 Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adrianson, Lillemor</td>
<td>Abrams, Zsuzsanna Ittzes</td>
</tr>
<tr>
<td>Baym, Nancy K.</td>
<td>Li, Qing</td>
</tr>
<tr>
<td>Collis, Betty</td>
<td>MacDonald, Lucy</td>
</tr>
<tr>
<td>Dyrli, Odvard Egil</td>
<td>Baron, Naomi S</td>
</tr>
<tr>
<td>Harris, Judith B.</td>
<td>Belz, Julie A</td>
</tr>
<tr>
<td>Hilz, Starr Roxanne</td>
<td>Benbunan-Fich, Raquel</td>
</tr>
<tr>
<td>Lea, Martin</td>
<td>Caverly, David C.</td>
</tr>
<tr>
<td>Mantovani, Giuseppe</td>
<td>Fahy, Patrick J</td>
</tr>
<tr>
<td>Matheson, Kimberly</td>
<td>Flanagan, Andrew J.</td>
</tr>
<tr>
<td>McMurd, George</td>
<td>Gu, guen, Nicolas</td>
</tr>
<tr>
<td>Olaniran, Bolanle A</td>
<td>Hampton, Keith N.</td>
</tr>
<tr>
<td>Phillips, Gerald M.</td>
<td>Haythornthwaite, Caroline</td>
</tr>
<tr>
<td>Rice, Ronald E.</td>
<td>Herring, Susan C</td>
</tr>
<tr>
<td>Riel, Margaret</td>
<td>Johnson, E. Marcia</td>
</tr>
<tr>
<td>Rojo, Alejandra</td>
<td>Kling, Rob</td>
</tr>
<tr>
<td>Schrum, Lynne</td>
<td>Kock, Ned</td>
</tr>
<tr>
<td>Snyder, Herbert</td>
<td>Lee, Lina</td>
</tr>
<tr>
<td>Valacich, Joseph S.</td>
<td>Selwyn, Neil</td>
</tr>
<tr>
<td>Walther, Joseph B.</td>
<td>Trentin, Guglielmo</td>
</tr>
<tr>
<td>Weinberg, Nancy</td>
<td>Tu, Chih-Hsiung</td>
</tr>
<tr>
<td>Zack, Michael H</td>
<td>Vrooman, Steven S.</td>
</tr>
<tr>
<td></td>
<td>Walther, Joseph B.</td>
</tr>
<tr>
<td></td>
<td>Warnick, Barbara</td>
</tr>
<tr>
<td></td>
<td>Wellman, Barry</td>
</tr>
<tr>
<td></td>
<td>Wilson, E. Vance</td>
</tr>
<tr>
<td></td>
<td>Wolfe, Joanna</td>
</tr>
</tbody>
</table>
**Table 4. Common core terms for CMC**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer-Mediated Communication</td>
<td>210%</td>
<td>960</td>
<td>1</td>
<td>457</td>
<td>1</td>
</tr>
<tr>
<td>Higher Education</td>
<td>250%</td>
<td>510</td>
<td>2</td>
<td>204</td>
<td>2</td>
</tr>
<tr>
<td>Internet</td>
<td>206%</td>
<td>287</td>
<td>3</td>
<td>139</td>
<td>3</td>
</tr>
<tr>
<td>Distance Education</td>
<td>322%</td>
<td>245</td>
<td>4</td>
<td>76</td>
<td>8</td>
</tr>
<tr>
<td>Computer-Assisted Instruction</td>
<td>198%</td>
<td>212</td>
<td>5</td>
<td>107</td>
<td>6</td>
</tr>
<tr>
<td>Foreign Countries</td>
<td>268%</td>
<td>153</td>
<td>7</td>
<td>57</td>
<td>14</td>
</tr>
<tr>
<td>Educational Technology</td>
<td>317%</td>
<td>152</td>
<td>8</td>
<td>48</td>
<td>18</td>
</tr>
<tr>
<td>Computer Uses in Education</td>
<td>219%</td>
<td>151</td>
<td>9</td>
<td>69</td>
<td>10</td>
</tr>
<tr>
<td>Online Systems</td>
<td>332%</td>
<td>146</td>
<td>10</td>
<td>44</td>
<td>19</td>
</tr>
<tr>
<td>Electronic Mail</td>
<td>78%</td>
<td>107</td>
<td>12</td>
<td>137</td>
<td>4</td>
</tr>
<tr>
<td>Information Technology</td>
<td>139%</td>
<td>71</td>
<td>19</td>
<td>51</td>
<td>17</td>
</tr>
<tr>
<td>Computer Networks</td>
<td>51%</td>
<td>70</td>
<td>20</td>
<td>136</td>
<td>5</td>
</tr>
<tr>
<td>Teleconferencing</td>
<td>111%</td>
<td>69</td>
<td>21</td>
<td>62</td>
<td>12</td>
</tr>
</tbody>
</table>

**Table 5. Unique core terms for CMC**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>World Wide Web</td>
<td>175</td>
<td>6</td>
<td>Telecommunications</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>Telematics</td>
<td>142</td>
<td>11</td>
<td>Information Networks</td>
<td>72</td>
<td>9</td>
</tr>
<tr>
<td>Second Language Learning</td>
<td>93</td>
<td>13</td>
<td>Interpersonal Communication</td>
<td>67</td>
<td>11</td>
</tr>
<tr>
<td>Student Attitudes</td>
<td>90</td>
<td>14</td>
<td>Computer Applications</td>
<td>57</td>
<td>13</td>
</tr>
<tr>
<td>Teaching Methods</td>
<td>83</td>
<td>15</td>
<td>Adulthood</td>
<td>53</td>
<td>15</td>
</tr>
<tr>
<td>Second Language Instruction</td>
<td>82</td>
<td>16</td>
<td>Computers</td>
<td>52</td>
<td>16</td>
</tr>
<tr>
<td>Online Courses</td>
<td>81</td>
<td>17</td>
<td>Experimental Theoretical</td>
<td>42</td>
<td>20</td>
</tr>
<tr>
<td>Interaction</td>
<td>78</td>
<td>18</td>
<td>Group Dynamics</td>
<td>42</td>
<td>21</td>
</tr>
</tbody>
</table>

**Table 6. Survey of research terms**

<table>
<thead>
<tr>
<th>Concept</th>
<th>1997</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case Studies</td>
<td>29</td>
<td>37</td>
</tr>
<tr>
<td>Communication Research</td>
<td>38</td>
<td>36</td>
</tr>
<tr>
<td>Comparative Analysis</td>
<td>21</td>
<td>32</td>
</tr>
<tr>
<td>Content Analysis</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>Discourse Analysis</td>
<td>14</td>
<td>28</td>
</tr>
<tr>
<td>Educational Research</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>Evaluation Criteria</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Evaluation Methods</td>
<td>7</td>
<td>31</td>
</tr>
<tr>
<td>Literature Reviews</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>Pilot Projects</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Research</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Research Methodology</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>Surveys</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>Use Studies</td>
<td>21</td>
<td>14</td>
</tr>
</tbody>
</table>
The more exhaustive 2005 analysis was used to identify research-affiliated terms to be inlaid within respective conceptual maps of the two studies. The research-affiliated concepts are listed for both the 1997 and 2005 research studies in Table 6. The highest number of scholarly articles was connected with the term communication research.

The top-ranked term was communication research for the 1997 survey and case studies for the 2005 survey. As would be expected, most concepts were identified at higher levels in the larger, more recent study. However, there were some exceptions to this trend. These included communication research, research methodology, and use studies.

**Conceptual Map of CMC Research Methodologies**

Research question 4 was explored in two ways. First was the identification of prominent problematisations within each of the previously identified research affiliated terms. This was done by setting a somewhat arbitrary four-occurrence threshold that had to be surpassed. What follows is a more lexical description of those relationships. Clearly, terms that are more numerous are not necessarily strongly related, and terms strongly related are not necessarily more numerous. Below are topical areas that were most dominant when the respective research problematisations were used as filters. Because of the magnitude of higher education occurrences, it was included in almost all categories.

**Communication research.** Both studies included higher education, Internet, communication behavior, and group dynamics as areas of foci. Additionally, the 2005 study included interpersonal relationships, student attitudes, and World Wide Web. The 1997 conceptual set was relatively extensive including the following terms:

- Adulthood
- Communication

- Communication Social Aspects
- Computer Applications
- Computer Assisted Instruction
- CMC Systems
- Computer Networks
- Computers
- Decision Making
- Electronic Mail
- Group Decision Making
- Interpersonal Relations
- Organizational Communication
- Social Interaction
- Telecommunications

**Case studies.** Both studies included the indexical terms higher education, Internet, distance education, foreign countries, computer-assisted instruction, and electronic mail. The 2005 study also had World Wide Web, instructional effectiveness, online systems, information technology, college students, literacy, and teacher role. The 1997 study was more writing oriented with computer networks, teaching methods, writing research, student attitudes, collaborative writing, technical writing, and writing instruction.

**Evaluation methods.** Only higher education and computer-assisted instruction appeared prominently in both studies when evaluations methods was used as a filter. However, no other terms, except for the previously mentioned higher education, reached a level of predominance in the 1997 study. The 2005 conceptual set was relatively extensive including the following terms:

- Collaborative learning
- Distance education
- Foreign countries
- Instructional effectiveness
- Interaction
- Internet
- Online systems
- Research methodology
- Student attitudes
- Student evaluation
Comparative analysis. Dominant concepts in both studies included higher education, distance education, and computer-assisted instruction. The 2005 study also had face-to-face communication, interaction, instructional effectiveness, interpersonal communication, nontraditional education, second-language instruction, and second-language learning. The 1997 study had more dominant terms including computer networks, futures of society, teleconferencing, electronic mail, and experimental theoretical.

Discourse analysis. Higher education and interpersonal communication was the only dominant concept in both studies. The 2005 dominant concepts included communication behavior, Internet, language research, second-language instruction, second-language learning, sex differences, and written language. Only two additional terms were dominant in the 1997 study. These were computer-assisted instruction and computer networks.

Educational research. The term educational research did not have sufficient mentions to be considered in the dominant term analysis for the 1997 study. The dominant terms for the 2005 study included higher education, computer-assisted instruction, cooperative learning, distance education, educational technology, Internet, nontraditional education, online systems, and World Wide Web.

Surveys. There were no dominant terms included in both studies other than higher education. The 2005 study featured student attitudes, distance education, and World Wide Web, while the 1997 study featured Internet, computer networks, electronic mail, teleconferencing, information networks, and scholarly communication.

Literature reviews. Distance education was a dominant term in both studies concerning literature reviews. The 2005 study also included community and information technology. The 1997 study additionally had higher education, computer networks, electronic mail, and teleconferencing.

Research. When used as a filter, the singular term, research, did not have sufficient mentions of any terminology to be included in this analysis.

Use studies. Higher education, Internet, and electronic mail were dominant in both studies. Additionally, the 2005 study featured the term World Wide Web. The 1997 study’s dominant concepts were more ranging, including communication thought transfer, computer networks, foreign countries, online systems, tables data, telecommunications, and teleconferencing.

Content analysis. Only higher education was identified for both studies. The only additional dominant term for the 2005 study was Internet. Terms for the 1997 study included electronic mail, interpersonal communication, research needs, and tables data.

Pilot project. There were no dominant mentions of terminology in both studies except for higher education. Additionally, the 2005 study included foreign countries and educational technology. The 1997 study had electronic mail.

Evaluation criteria. Evaluation criteria did not have sufficient enough mentions for dominant term analysis in the 1997 study. The dominant affiliated terms with the 2005 study included higher education, computer-assisted instruction, distance education, Internet, and online systems.

Research methodology. There was no overlap in dominant terms. The 2005 study had distance education and interaction, and the 1997 study had higher education, computer networks, and electronic mail.

The second way research question 4 was explored was to examine strength of affiliation through cluster analysis. This was thought to provide some clues as to which core terms were most closely related to particular research problematisations. Some of the linkages were clearly methodological (Table 7). However, not all terms
were included in both analyses due to the number of occurrences when research term filters were applied. Table 7 identifies clusters of research and affiliated co-occurring terms that appeared in at least one of the studies. Clustered terms tend to be conceptual sets that journal articles focus upon.

Aside from terms that were affiliated across both studies, there were terms that only appeared

Table 7. Longitudinal research term comparison

<table>
<thead>
<tr>
<th>2005 - research clusters</th>
<th>1997- research clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case studies</strong></td>
<td><strong>Case studies</strong></td>
</tr>
<tr>
<td>Case studies</td>
<td>group discussion</td>
</tr>
<tr>
<td>Literacy</td>
<td>graduate study</td>
</tr>
<tr>
<td>computer literacy</td>
<td>writing instruction</td>
</tr>
<tr>
<td>Tutoring</td>
<td></td>
</tr>
<tr>
<td><strong>Communication research</strong></td>
<td><strong>Communication research</strong></td>
</tr>
<tr>
<td>communication behavior</td>
<td>Organizational communication</td>
</tr>
<tr>
<td><strong>Discourse analysis</strong></td>
<td>interpersonal communication</td>
</tr>
<tr>
<td>classroom communication</td>
<td>Computers</td>
</tr>
<tr>
<td><strong>Comparative analysis</strong></td>
<td><strong>Discourse analysis</strong></td>
</tr>
<tr>
<td>face to face communication</td>
<td>microcomputers</td>
</tr>
<tr>
<td>interpersonal communication</td>
<td>man machine systems</td>
</tr>
<tr>
<td><strong>Content analysis</strong></td>
<td><strong>Research methodology</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Research methodology</strong></td>
<td><strong>Evaluation methods</strong></td>
</tr>
<tr>
<td><strong>Evaluation methods</strong></td>
<td>learner controlled instruction</td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td><strong>Content analysis</strong></td>
</tr>
<tr>
<td>electronic mail</td>
<td>research needs</td>
</tr>
<tr>
<td>ethics</td>
<td></td>
</tr>
<tr>
<td><strong>Use studies</strong></td>
<td><strong>Survey</strong></td>
</tr>
<tr>
<td>computer use</td>
<td>Teleconferencing</td>
</tr>
<tr>
<td>undergraduates</td>
<td>foreign countries</td>
</tr>
<tr>
<td>technology utilization</td>
<td></td>
</tr>
<tr>
<td><strong>Literature reviews</strong></td>
<td><strong>Use studies</strong></td>
</tr>
<tr>
<td>Community</td>
<td><strong>Literature reviews</strong></td>
</tr>
<tr>
<td></td>
<td>tables data</td>
</tr>
<tr>
<td></td>
<td>Questionnaires.</td>
</tr>
</tbody>
</table>

*=not enough for 1997 analysis
in the 2005 analysis. These relationship clusters included:

- **Educational Research**: Distance education, nontraditional education, computer-assisted instruction, and educational practices
- **Evaluation Criteria**: Feedback, course content, and course development
- **Pilot Project**: Information services, library services, reference services, user needs, information, and electronic libraries
- **Research**: Communication

### DISCUSSION

This chapter analyzed “what is meant by computer-mediated communication research?” While a number of suitable answers exist for this question, it chose to let the “field” define itself as it exists in the extant literature concerning business, education, psychology, and the social sciences. In this regard, almost 40 years of archival data was surveyed. Wallace’s study of 1997 and prior literature was combined with this 2005 examination to offer a look at how CMC research is viewed through a database analytic lens. Because of the interdisciplinary nature of computer-mediated communication (CMC), a parallel analysis of multiple databases from different perspectives was used (Ingwersen & Christensen, 1997; McLaughlin, 1994; Spasser, 1997). It should be noted that as database sophistication increases regarding education, business, and social life in general, this kind of lens may become more crucial.

The descriptive nature of this study necessarily dictated a balance between rigor and latitude. Several limitations must be considered in this respect. These include generalization, design limitations, and theoretical assumptions. First, operationalization of the domain should clearly impact the use of the findings. Therefore, results from this analysis do not claim to identify characteristics of CMC beyond the domain examined. Second, the selection of cluster analysis over other classifying procedures has a number of inherent limitations. This was considered a necessary condition given the emerging nature of the field. Emergence (or lack of emergence) of clusters will have to be weighed against the nonrandom nature of their partitioning (Aldenderfer & Blashfield, 1984). Third, the design was somewhat restricted in order to have a longitudinal comparison. The above limitations were addressed in the traditional way for descriptive studies. Detailed explanations, including supporting literature, were provided for choices made.

### CONCLUSION

This survey of computer-mediated communication literature revealed three interesting trends. The first trend is a paradoxical turbulence and stability common in literature surveys. This pattern was somewhat present for articles, journals, authors, concepts, and research affiliations. The total articles doubled while the number of core journals remained relatively constant. Both the overall production of articles and the production of core journals increased by 217%. Despite this increase, the number of core journals producing those articles only advanced by 3, from 22 to 25.

The total number of core authors producing more than two articles also had relatively little growth, while there was a virtual turnover in actual author names. Joe Walther was the only author to emerge in both surveys. Core authorship increased by an anemic 30%, from 21 to 31. When considering total articles produced, core author production actually shrank from 13% in the 1997 survey, to 8.5% in the 2005 survey. The top 21 terms in both studies accounted for 30% of the total indexical terms. Eight of those terms were not reciprocal. Most of the unique terms can be attributed to shifts in the more turbulent bottom half of the distribution. However,
telecommunications and information networks were ranked seventh and ninth, respectively, in the 1997 survey, and were not prominent in the 2005 study. One interesting note is that while “World Wide Web” was not a core terminology in the 1997 study, it was identified as the number one emerging concept. In the 2005 survey, it was firmly established as sixth in overall mentions. Communication research and discourse analysis were the only research affiliated terms that fell within the top 21 terms of both surveys.

The majority of the previously identified research terminologies emerged in both studies, but were not related to the same issues. Nine out of the 14 research affiliated terms emerged in both surveys. Serendipitously, the only two terms that were related to each other were also research affiliated. These were content analysis and research methodology. The terms educational research, literature reviews, research, pilot projects, and evaluation criteria were exclusive in that they did not overlap with any of the other dominant research terms.

The second trend is that literature tended to have an online and higher education focus. Five hundred and ten articles were higher-education related. Internet was the second most populous descriptor, with 287 articles. Furthermore, World Wide Web had 176 articles, and online systems had the highest percentage increase with 146 articles. Besides articles, both surveys indicated in excess of 70% of the journals were related to education. Core journals were also highly education and Internet related.

Higher education was also connected with almost all research terminology for both surveys. The only exceptions were the terms research, research methodology, and literature review for the 2005 survey. The 1997 study had higher education absent for educational research, evaluation criteria, and the singular term, research. While the 1997 survey results could be attributed to low production numbers, this was not the case with the 2005 results. Other problematisations with similar production numbers managed to have at least four articles with a higher education affiliation.

The third trend emerged through a post hoc analysis of the 2005 dataset that was not conducted in the 1997 study. Journals were examined to see if they affiliated with any of the dominant research terminology. Most journals were affiliated with a single prominent research term. Eight were affiliated with two terms, and only the journal Internet Research involved three terms that included comparative analysis, evaluation criteria, and use studies. The appendix describes the research affiliation as indicated by key word term, journal that it was published in, and the count of articles in a journal concerning a particular research term.

FUTURE RESEARCH

Clearly, this study detailed CMC research as an area that tended to be education and “Internet” affiliated. Furthermore, the computer-mediated communication literature prominently used a number of textual analysis techniques, such as content and discourse analysis. Noticeably absent were articles focused on possible experimental techniques, ethnographies, and focus groups. This does not mean that these were not tools used in the literature, merely that they were not the focus of the more prominently presented research articles (Schneider & Foot, 2004). Surveys also had a surprisingly diminutive presence. Certainly, there are both specific and extensive treatments of online survey methodology (e.g., Andrews, Nonnecke, & Preece 2003; Katz & Rice, 2002). However, the current examination suggests a need for this and other volumes to specifically localize methodological issues relevant to computer-mediated and online communication.

With the exception of Joe Walther, core authors had completely overturned. While Nicholls concedes the robustness of the straight count, an exhaustive identification of authorship might more
fully answer questions concerning ranking issues, shifting authorships, and omissions (Cronin & Overfelt, 1994; Nicholls, 1989).

One general consideration is that indexed literature does not always reflect the centrality of artifacts and producers. The current analysis focused on work being produced without discrimination in regard to usage. Identification of CMC’s literature usage patterns might provide complementary information to the current study. Furthermore, it would help shed light on whether theoretical or pragmatic issues are driving CMC’s scholarly communication.

Another important consideration is that it is unlikely that people engaged in CMC are overwhelmingly involved in education-related phenomena (Katz & Rice 2002; Papacharissi & Rubin, 2000). If the study of CMC is to more accurately reflect current usage patterns, it should be aggressively broadened in other areas. Conceptual linkages exposed by this analysis are formative at best. Future studies should rigorously examine both the linkages and the conclusions in regard to their longevity and stability.

However, while conceding a number of caveats, this study does identify component features of past and more recent computer-mediated communication research. These resources are readily available online through many library services and outside vendors. Therefore, CMC and “online” researchers can use these features to identify research redundancies, and opportunities to consider, and hopefully more efficiently advance, the scholarship of the field.

REFERENCES


**KEY TERMS**

**Artifacts:** Artifacts are any number of forms of scholarly communication. Conceptually, they could range from working papers to books.

**Bibliographic Coupling:** Bibliographic coupling is where two documents each have citations to one or more of the same publication, but do not have to necessarily cite each other.

**Bibliometrics:** The mathematical and statistical analysis of patterns that arise in the publication and use of documents.

**Bradford Partitions:** These partitions are used in library and information science to establish core journals. The process ranks journals from most to least prolific in terms of number of articles produced concerning a subject. They are then divided into three or more “zones” that have roughly the same number of articles. The zone that has the least journals is used to identify the “core” journals for a subject area.

**CMC:** An acronym standing for computer-mediated communication or computer mediated communication.

**Co-Citation Analyses:** The analysis of journals, articles, or authors that are cited together in an article or articles.

**Concepts:** Concepts are terms or words used by the producers themselves. This could include words in titles and text, or assigned terminology such as key words.

**Core Authors:** Core authors are generally established through a well-articulated benchmark that exceeds two publications.

**Core Concepts:** Core concepts are the relatively few concepts that account for a large amount of problematisations under study.

**Core Journals:** Core journals are generally established through a well-established benchmark such as Bradford partitions.

**Co-Word Analyses:** The analyses of the co-occurrence of two or more words in one document or in different documents.
### APPENDIX: 1997-2005

<table>
<thead>
<tr>
<th>Journal</th>
<th>Count</th>
<th>Journal</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case studies</td>
<td></td>
<td>Evaluation methods</td>
<td></td>
</tr>
<tr>
<td>Computers and Composition</td>
<td>2</td>
<td>American Journal of Distance Education</td>
<td>2</td>
</tr>
<tr>
<td>Educational Media International</td>
<td>2</td>
<td>Instructional Science</td>
<td>2</td>
</tr>
<tr>
<td>Internet and Higher Education</td>
<td>2</td>
<td>Quarterly Review of Distance Education</td>
<td>2</td>
</tr>
<tr>
<td>Journal of Computer Assisted Learning</td>
<td>2</td>
<td>Educational Technology &amp; Society</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Computers &amp; Education</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Journal of Educational</td>
<td>2</td>
</tr>
<tr>
<td>Communication research</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Communication Research</td>
<td>3</td>
<td>Computing Research</td>
<td>2</td>
</tr>
<tr>
<td>Computers and Composition</td>
<td>3</td>
<td>Distance Education</td>
<td>2</td>
</tr>
<tr>
<td>Public Relations Review</td>
<td>2</td>
<td>Performance Improvement</td>
<td>2</td>
</tr>
<tr>
<td>International Review of Research in Open and Distance Learning</td>
<td>2</td>
<td>Literature reviews</td>
<td></td>
</tr>
<tr>
<td>Business Communication Quarterly</td>
<td>2</td>
<td>ARIST</td>
<td>6</td>
</tr>
<tr>
<td>Journal of Computer-Mediated Communication</td>
<td>2</td>
<td>Evaluation methods</td>
<td></td>
</tr>
<tr>
<td>Written Communication</td>
<td>2</td>
<td>Internet Research</td>
<td>2</td>
</tr>
<tr>
<td>Comparative analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet Research</td>
<td>3</td>
<td>Educational Media International</td>
<td>4</td>
</tr>
<tr>
<td>Computers &amp; Education</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content analysis</td>
<td></td>
<td>Journal of Language &amp; Social Psychology</td>
<td>2</td>
</tr>
<tr>
<td>Educational Computing Research</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructional Science</td>
<td>2</td>
<td>Survey</td>
<td></td>
</tr>
<tr>
<td>Discourse analysis</td>
<td></td>
<td>Educational Media International</td>
<td>2</td>
</tr>
<tr>
<td>Computers and Composition</td>
<td>6</td>
<td>American Journal of Distance Education</td>
<td>2</td>
</tr>
<tr>
<td>CALICO Journal</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research on Language and Social Interaction</td>
<td>3</td>
<td>Internet Research</td>
<td>5</td>
</tr>
<tr>
<td>Written Communication</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Educational research</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>British Journal of Educational Technology</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet and Higher Education</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal of Adolescent &amp; Adult Literacy</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quarterly Review of Distance Education</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community College Journal</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This work was previously published in the Handbook of Research on Electronic Surveys and Measurements, edited by R. Reynolds, R. Woods, and J. Baker, pp. 207-222, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 1.24
Computer-Mediated Communication in Virtual Learning Communities

Lisa Link
Flensburg University of Applied Sciences, Germany

Daniela Wagner
University of Hildesheim, Germany

THE MEANING OF CMC

Computer-mediated communication (CMC) can be simply defined as “communication that takes place between human beings via the instrumentality of computers” (Herring, 1996, p. 1) and the technology used in CMC as “media that facilitate the exchange of semantic content, transmitted through telecommunication networks, processed through one or more computers, between individuals and among groups” (Rice, 1984, p. 438). As a concomitant result of the widespread use of computers and the Internet in the curriculum of higher education as well as the inclusion of various forms of online learning, CMC has become a common communication modality in teaching and learning contexts. Although all forms of CMC can be meaningfully integrated into learning contexts, the majority of CMC at present is nevertheless text based and hence the focus of this contribution.

CMC plays a pivotal role in the context of e-learning models in higher education that offer students new forms of learning that would not be possible in a traditional classroom environment. Cooperative and collaborative learning, problem-oriented and project-based learning, as well as authentic learning in which real-world scenarios are simulated, are some examples. In the context of these types of e-learning models, virtual learning communities as well as smaller virtual teams are often an essential component. The learning objectives of these models are manifold, but they generally include offering students the opportunity to acquire, practice and enhance important key competencies required in their future professional work. Hence, communication and media skills as well as the ability to work cooperatively
in a virtual team are just as important as the actual subject matter of the course itself.

**CHARACTERISTICS OF CMC**

CMC can be classified into two major groups: asynchronous and synchronous CMC. The main difference between these two types is temporal: asynchronous CMC is time independent, that is, it does not require that the communication partners be simultaneously online, whereas synchronous CMC takes place in real time or quasi real time, requiring the telepresence of the communication partners. E-mail, mailing lists, and discussion forums are examples of asynchronous forms. Chat rooms and shared whiteboards represent synchronous forms of CMC.

A further classification of CMC is whether it represents a one-to-one (1:1), one-to-many (1:n) or many-to-many (n:n) communication form. Depending on their use, the different types of CMC can fall into more than one category, for example, e-mail and chat can represent both 1:1 and n:n communication. A topic of interest in this context is the double function CMC can have: It can be used for individual communication, but also for mass communication. This goes along with a double function that is very interesting in a learning setting, for example, in higher education. E-mail messages and discussion forum postings can simultaneously fulfill two successive functions: (1) interpersonal communication between two or more participants and subsequently (2) serve as an information pool for other participants. Chats that have a protocol option can also be used as an information pool for passive students. Fritsch (1998) coined the term *witness learning* to describe the indirect learning possibilities of learners who do not actively take part in interactions, but learn from witnessing the interactions of others. In virtual learning environments, participants have ranked “witnessing” (i.e., reading) the interactions of others high among the things they have learned from (Fritsch, 1998; Link, 2002).

As text-based CMC is not realized face-to-face but mediated via computers, the communication partners cannot directly experience nonverbal signals (e.g., facial expressions or gestures) or paraverbal signals (e.g., voice volume and tone). In practice, nonverbal and paraverbal signals are often compensated for by emoticons, inflectives, and other expressions created by entering characters on the keyboard.

Many studies, particularly those from a linguistic perspective, have investigated the style participants use in communicating. The most often noted phenomenon is no doubt the use of colloquial language in chat, e-mail, and even in discussion forums. This phenomenon is often accompanied by unconventional orthography, spelling and grammar errors, as well as slang expressions. Furthermore, the communication is influenced by a register embracing word creations taken from the language of computer hackers. If the model of Koch and Österreicher (1994) is applied to CMC, it becomes apparent that the linguistic characteristics of this type of written communication cannot simply be classified as totally oral or totally written: Here it is helpful to differentiate between the concept and medium of spoken and written language. Considering the communication style of CMC, it must be regarded as a hybrid language variety, displaying characteristics of both spoken and written language. As its usage norms are becoming conventionalised, some authors venture to propose that this communication style represents a new type of language which they refer to as netspeak or cybertalk, which cannot be compared to conventional communication (e.g., Crystal, 2001).

Studies with a focus on work-flow analysis often examine CMC as to its impact on the effectiveness of communication in professional contexts as well as on the dependency between communication style and the roles of the commu-
Computer-Mediated Communication in Virtual Learning Communities

mediation partners (e.g., communication between team colleagues vs. communication between employee and employer).

With regard to the literature and research reports published in the last 10 years, CMC plays an important role in nearly every community phenomenon. In social contexts, the style people use to communicate influences to a high degree the social relations and team dynamics.

**CMC THEORIES**

For the effective use of CMC in educational contexts, a variety of computer-mediated communication theories can provide insights into selecting appropriate CMC tools as well as understanding their limitations. Prevalent theories can be categorized into three large groups (Döring, 2003, p. 128):

1. Media choice (e.g., media richness theory: Daft & Lengel, 1984, 1986; Rice, 1992; social presence: Short, Williams, & Christie, 1976; Rice, 1993)
2. Media characteristics (e.g., cues-filtered-out model: Culnan & Markus, 1987; reduced social cues: Sproull & Kiesler, 1988)
3. Mediated communication behaviour (e.g., social information processing: Walther, 1992; Fulk, Schmitz, & Steinfeld, 1990).

Media choice theories focus on cognisant decision making about which medium to use for specific communicative tasks. The media richness theory (Daft & Lengel, 1984, 1986; Rice, 1992) is by far the most predominant theory of media choice. This theory defines a medium’s richness in terms of its capacity to carry information, feedback, channel, source, and language. Placing face-to-face communication at the richest end of the spectrum and numerical computer documents at the leanest, various communication media are then placed along this scale. In the media richness model of Reichwald, Möslin, Sachenbacher, & Englberger (2000, p. 57), which includes most forms of CMC, the spectrum from rich to lean media range from (1) face-to-face dialog, (2) video conference, (3) telephone/telephone conference, (4) voice mail, (5) computer conference, (6) telefax, (7) e-mail, (8) letter post and documentation. This theory contends that in organisations information is processed in order to reduce uncertainty or equivocality. Reducing uncertainty (e.g., communicating the postponement of a scheduled meeting) can be carried out best with lean media whereas reducing equivocality (e.g., a team meeting about a new project) requires a rich medium. Thus, the primary assertion of this theory is that the information richness of the task should correspond to the media richness of the medium.

To date, empirical studies have not been able to validate this theory.

Media characteristic theories examine the impact that media characteristics of CMC have on the communication process. The starting point is that CMC is typically text based. Accordingly, this results in reducing communication channels to text only, which in turn filters out physical and social cues. The cues-filtered-out approach (Culnan & Markus, 1987) propounds that since text-based CMC cannot communicate nonverbal or paraverbal information, very little social and sociodemographic information about the communication participants is transmitted. Additional research has found that the lack of social cues can be liberating, for example, in decreasing inhibition (Kiesler, Siegal, & McGuire, 1984).

Mediated communication behaviour theories focus on how participants behave during CMC, whether they exchange social information, form relational bonds, or use special Internet language. Social information processing (Walther, 1992) maintains that with time, CMC participants will develop communication skills to compensate for limitations of the medium. In contrast to the other
CMC theories, social information procession views the Internet as a new social environment in which over time communication partners find new ways in which to integrate nonverbal social information into their online behaviour.

**THE ROLE OF CMC IN VIRTUAL LEARNING COMMUNITIES IN HIGHER EDUCATION**

In view of the substantial use of CMC in many professional fields, especially in those dealing with research and development projects and distributed teamwork, higher education should prepare students to work effectively with CMC tools and familiarise them with the characteristics of CMC situations as well as with the impact these can have on work efficiency.

Even if many people know how to use an e-mail program or a chat tool or how to place a message in a Web-based bulletin board, this does not mean that they are automatically aware of the advantages and disadvantages of CMC. In this respect, there is still a significant lack of information and need for training which should be addressed in higher education. The boom in e-learning research projects has afforded most institutes of higher education the technical infrastructure necessary to offer students learning opportunities in virtual learning communities.

The following concept is just one example of how students can achieve the above-mentioned key competencies in virtual learning communities in higher education. The concept has been developed and successfully implemented in the project MEUM, a joint pilot project of the Flensburg University of Applied Sciences and the University of Hildesheim in Germany. Both institutes offer specialised degree programmes in translation and have teamed up to design and implement a distributed telelearning programme in which students can learn about and gain firsthand practical experience in Internet-based cooperative work in the field of translation. The students work in virtual teams on a concrete translation project. Here, we define a virtual team as a geographically distributed team, whose members communicate mainly via CMC. A detailed description of the MEUM project can be found in Wagner and Link (2003).

The blended learning courses offered in the context of MEUM are characterised by three phases:

1. Orientation phase in which the students receive initial information about the course didactics, objectives, and schedule and are organised into teams. This phase also includes an introduction to the Internet-based learning platform in which both learning modules and CMC tools (e-mail, chat, and discussion forums) are integrated;
2. Teamwork phase during which the virtual teams work autonomously to prepare and carry out their translation job; and
3. Team assessment and reflection phase in which the students reflect on their team and work process as well as on the impact the CMC and the overall virtuality had on their work.

In these courses, students are offered information on the theory and characteristics of CMC, for example, in Web-based learning modules. The virtual teams are free to choose which communication tools to use for the various tasks and phases of their virtual teamwork. This has resulted in all CMC tools being utilised. This combination of theory, task-based teamwork, and joint reflection phase has been rapidly accepted by the students and the reflection phase, in particular, is seen by them as a vital component of this concept.

For a successful transfer of this concept to other courses, it is important to consider the competencies required of students and instructors in a virtual learning community.
Computer-Mediated Communication in Virtual Learning Communities

Competencies of the instructors: Virtual learning communities require instructors to take on the supportive role of mentors or tutors in the background, functioning, for example, as consultants on demand. In general, they prepare the introduction of the virtual community and during the course monitor the activities of the communities from behind the scenes. In a preparatory phase, the instructor decides whether face-to-face meetings should be integrated, and if so, whether the team members should meet their team colleagues at the start. Prior to the course, students should be asked to give information about their media and professional skills as well as their prior experience with (virtual) teamwork. Instructors must also decide how teams should be formed: should students be given the task of building their own teams or is it advantageous to compose comparable teams according to student competency profiles?

As our own experience in virtual learning communities has shown, it is, on the one hand, very important to explain the role of the instructors in detail to the students and, on the other hand, for instructors to remain aware of their own role as tutors or mentors. By monitoring the work phase in the learning communities, instructors can decide when it might be necessary to intervene in the teamwork. This presupposes a didactic model in which the tutor/instructor has a consulting function in the background.

Competencies of the students: The following can be considered the requisite competencies for members in a virtual learning community as well as the fundamental skills and abilities that students should have at the end of the course.

The members of a virtual learning community are aware of the characteristics of and differences between the CMC tools chat, e-mail, and discussion forums. They have knowledge of the standard communication conventions for CMC (e.g., netiquette and chatiquette). The members are familiar with emoticons and other strategies for compensating for the lack of nonverbal and paraverbal signals in CMC, as well with moderation techniques, for example, in chats with more than two participants. The virtual community members know the importance of feedback for asynchronous communication, for example, a confirmation that an e-mail has arrived or a response to a note in a discussion forum to encourage the communication. They understand the role of small talk in CMC and the difficulties that go along with the mixture of task-oriented and relationship-oriented communication. It is important that community members recognise when communication results in misinterpretation and are able to identify the problems that CMC can produce in the learning community.

In the wake of globalisation, CMC and distributed virtual teams will become more and more commonplace in professional and educational contexts. At present we are still in an experimental phase that calls for more empirical studies into the efficient and effective use of these phenomena.

At the same time, future development of virtual learning communities should be extended across cultural and discipline boundaries. A new challenge in virtual learning communities will thus be intercultural and interdisciplinary cooperation.

CONCLUSION

In this article we presented a definition and classification of CMC. CMC is a phenomenon that is studied by a wide variety of disciplines: linguists, social psychologists, and computer scientists have proposed approaches to help understand the particularities and impacts of CMC. In addition to a discussion of selected CMC theories, we presented a sensibilization concept for CMC in higher education with the aim of helping students and instructors attain the key competencies required of members of a virtual (learning) community. The dissemination of distributed virtual
teamwork in professional fields requires new skill and competency requirements, which in turn must be addressed in educational training. Giving students the opportunity to learn how to utilise information and communication technology as well as to develop suitable strategies in CMC is a new challenge for higher education.

REFERENCES


**KEY TERMS**

**Blended Learning:** Learning design that combines various activities such as face-to-face meetings, Internet-based learning modules, and virtual learning communities.

**Chatiquette:** Standard rules of courtesy and correct behaviour for online chat.

**Computer-Mediated Communication (CMC):** Communication between humans using the computer as a medium.

**Emoticons:** A combination of punctuation marks and other special characters from the keyboard used to convey the tone of a computer-mediated communication message. For example, the combination :-) depicts smiling.

**Learning Platform:** Software systems that are used to deliver and support online teaching and learning. Learning platforms manage access to the platform and to learning materials and usually include various communication tools.

**Netiquette:** Standard rules of courtesy and correct behaviour on the Internet.

**Witness Learning:** A term coined by Dr. Helmut Fritsch, senior researcher at the FernUniversität in Hagen, Germany, that refers to the indirect learning possibilities of learners who do not actively take part in interactions, but learn from witnessing the interactions of others, for example, in online discussion forums.

This work was previously published in the Encyclopedia of Virtual Communities and Technologies, edited by S. Dasgupta, pp. 49-53, copyright 2006 by Information Science Reference; formerly known as Idea Group Reference (an imprint of IGI Global).
INTRODUCTION

Historically, computer security has its roots in the military domain with its hierarchical structures and clear and normative rules that are expected to be obeyed (Adams & Sasse, 1999). The technical expertise necessary to administer most security tools stems back to the time where security was the matter of trained system administrators and expert users. A considerable amount of money and expertise is invested by companies and institutions to set up and maintain powerful security infrastructures. However, in many cases, it is the user’s behavior that enables security breaches rather than shortcomings of the technology. This has led to the notion of the user as the weakest link in the chain (Schneier, 2000), implying that the user was to blame instead of technology. The engineer’s attitude toward the fallible human and the ignorance of the fact that technology’s primary goal was to serve human turned out to be hard to overcome (Sasse, Brostoff, & Weirich, 2001).

BACKGROUND

With the spreading of online work and networked collaboration, the economic damage caused by security-related problems has increased considerably (Sacha, Brostoff, & Sasse, 2000). Also, the increasing application of personal computers, personal networks, and mobile devices with their support of individual security configuration can be seen as one reason for the increasing problems with security (e.g., virus attacks from personal notebooks, leaks in the network due to personal wireless LANs, etc.) (Kent, 1997). During the past decade, the security research community has begun to acknowledge the importance of the human factor and has started to take research on
human-computer interaction into consideration. The attitude has changed from blaming the user as a source of error toward a more user-centered approach trying to persuade and convince the user that security is worth the effort (Ackerman, Cranor, & Reagle, 1999; Adams & Sasse, 1999; Markotten, 2002; Smetters & Grinter, 2002; Whitten & Tygar, 1999; Yee, 2002).

In the following section, current research results concerning the implications of user attitude and compliance toward security systems are introduced and discussed. In the subsequent three sections, security-related issues from the main application areas, such as authentication, email security, and system security, are discussed. Before the concluding remarks, an outlook on future challenges in the security of distributed context-aware computing environments is given.

USER ATTITUDE

The security of a system cannot be determined only by its technical aspects but also by the attitude of the users of such a system. Dourish et al. (2003) distinguish between theoretical security (e.g., what is technologically possible) and effective security (e.g., what is practically achievable). Theoretical security to their terms can be considered as the upper bound of effective security. In order to improve effective security, the everyday usage of security has to be improved. In two field studies, Weirich and Sasse (2001) and Dourish et al. (2003) explored users’ attitudes to security in working practice. The findings of both studies can be summarized under the following categories: perception of security, perception of threat, attitude toward security-related issues, and the social context of security.

Perception of security frequently is very inaccurate. Security mechanisms often are perceived as holistic tools that provide protection against threats, without any detailed knowledge about the actual scope. Therefore, specialized tools often are considered as insufficient, as they do not offer general protection. On the other hand, people might feel protected by a tool that does not address the relevant issue and thus remain unprotected (e.g., firewall protects against e-mail virus).

Perception of threats also reveals clear misconceptions. None of the users asked considered themselves as really endangered by attacks. As potential victims, other persons in their organization or other organizations were identified, such as leading personnel, people with important information, or high-profile institutions. Only a few of them realized the fact that they, even though not being the target, could be used as a stepping stone for an attack. The general attitude was that no one could do anything with the information on my computer or with my e-mails.

Potential attackers mainly were expected to be hackers or computer kids, with no explicit malevolent intentions but rather seeking fun. Notorious and disturbing but not really dangerous offenders, such as vandals, spammers, and marketers, were perceived as a frequent threat, while on the other hand, substantially dangerous attackers such as criminals were expected mainly in the context of online banking.

The attitude toward security technology was rather reserved. Generally, several studies reported three major types of attitudes toward security: privacy fundamentalists, privacy pragmatists, and privacy unconcerned (Ackerman et al., 1999). Users’ experiences played a considerable role in their attitude, as experienced users more often considered security as a hindrance and tried to circumvent it in a pragmatic fashion in order to reach their work objectives. Weirich and Sasse (2001) report that none of the users absolutely obeyed the prescribed rules, but all were convinced that they would do the best they could for security.

Additionally, users’ individual practices are often in disagreement with security technology.
People use legal statements in e-mail footers or cryptic e-mails, not giving explicit information but using contextual cues instead. In conjunction with such subsidiary methods and the fact that people often seem to switch to the telephone when talking about important things (Grinter & Palen, 2002) indicates the poor perception users have of security technology.

The feeling of futility was reported with respect to the need for constantly upgrading security mechanisms in a rather evolutionary struggle (i.e., if somebody really wants to break in, he or she will). As a result, personal accountability was not too high, as users believed that in a situation where someone misused his or her account, personal credibility would weigh more than computer-generated evidence, in spite of the fact that the fallibility of passwords is generally agreed.

The social context has been reported to play an important role in day-by-day security, as users are not permanently vigilant and aware of possible threats but rather considered with getting their work done. Therefore, it is no wonder that users try to delegate responsibility to technical systems (encryption, firewalls, etc.), colleagues and friends (the friend as expert), an organization (they know what they do), or institutions (the bank cares for secure transfers). Most people have a strong belief in the security of their company’s infrastructure. Delegation brings security out of the focus of the user and results in security unawareness, as security is not a part of the working procedure anymore.

Whenever no clear guidelines are available, people often base their practice on the judgments of others, making the system vulnerable to social engineering methods (Mitnick, Simon, & Wozniak, 2002). In some cases, collaboration appears to make it necessary or socially opportune to disclose one’s password to others for practical reasons, technical reasons, or as a consequence of social behavior, since sharing a secret can be interpreted as a sign of trust. Such sharing is a significant problem, as it is used in social engineering in order to obtain passwords and to gain access to systems.

Dourish et al. (2003) came to the conclusion that “where security research has typically focused on theoretical and technical capabilities and opportunities, for end users carrying out their work on computer systems, the problems are more prosaic” (p. 12). The authors make the following recommendations for the improvement of security mechanisms in the system and in the organizational context:

- Users should be able to access security settings easily and as an integral part of the actions, not in the separated fashion as it is today; therefore, security issues should be integrated in the development of applications (Brostoff & Sasse, 2001; Gerd tom Markotten, 2002).
- It is necessary that people can monitor and understand the potential consequences of their actions (Irvine & Levin, 2000) and that they understand the security mechanisms employed by the organization.
- Security should be embedded into working practice and organizational arrangement, and visible and accessible in everyday physical and social environment (Ackerman & Cranor, 1999).
- Security should be part of the positive values in an organization. So-called social marketing could be used to establish a security culture in a company.
- The personal responsibility and the danger of personal embarrassment could increase the feeling of personal liability.
- The importance of security-aware acting should be made clear by emphasizing the relevance to the organization’s reputation and financial dangers.
As has been shown, the design and implementation of security mechanisms are closely interlinked to the psychological and sociological aspects of the user’s attitude and compliance toward the system. Any security system is in danger of becoming inefficient or even obsolete if it fails to provide adequate support and motivate users for its proper usage. The following sections discuss these findings in the context of the main application domains of computer security.

AUTHENTICATION

Information technology extends our ability to communicate, to store and retrieve information, and to process information. With this technology comes the need to control access to its applications for reasons of privacy and confidentiality, national security, or auditing and billing, to name a few. Access control in an IT system typically involves the identification of a subject, his or her subsequent authentication, and, upon success, his or her authorization to the IT system.

The crucial authentication step generally is carried out based on something the subject knows, has, or is. By far the most widespread means of authentication is based on what a subject has (e.g., a key). Keys unlock doors and provide access to cars, apartments, and contents of a chest in the attic. Keys are genuinely usable—four-year-olds can handle them. In the world of IT, something the subject knows (e.g., a password or a secret personal identification number [PIN]) is the prominent mechanism.

The exclusiveness of access to an IT system protected by a password rests on the security of the password against guessing, leaving aside other technical means by which it may or may not be broken. From an information theoretic standpoint, a uniformly and randomly chosen sequence of letters and other symbols principally provides the greatest security. However, such a random sequence of unrelated symbols also is hard to remember, a relation that is rooted in the limitation of humans’ cognitive capabilities.

As a remedy, a variety of strategies were invented to construct passwords that humans can memorize more easily without substantially sacrificing the security of a password (e.g., passwords based on mnemonic phrases). For instance, Yan, et al. (2000) conducted a study with 400 students on the effect of three forms of advice on choosing passwords. They found, for example, that passwords based on pass phrases were remembered as easily as naively chosen passwords, while being as hard to crack as randomly chosen passwords. Insight into human cognition also has led to the investigation of alternatives such as cognitive passwords (Zviran & Haga, 1990), word associations (Smith, 1987), pass phrases (Spector & Ginzberg, 1994), images (Dhamija & Perrig, 2000), or pass faces (Brostoff & Sasse, 2000).

Authentication in public places, as is the case with automatic teller machines (ATM), has turned out to be vulnerable to attacks, where criminals obtain a user’s PIN by using cameras or other methods of observation in so-called shoulder-surfing attacks (Colville, 2003). In order to obscure the numbers entered by the user and thus hamper the recording of the necessary PIN, several techniques have been proposed (Hopper & Blum, 2001; Wilfong, 1997). Recently, Roth, et al. (2004) suggested variants of cognitive trapdoor games to protect users against shoulder-surfing attacks. In this approach, the buttons on a PIN pad are colored either black or white, and the user has to decide whether the number to be entered is in the black or white group. As the colors are changing randomly, the user has to enter the same number three to four times to complete an input. By blurring the response set with so-called shadows, this method can be made resistant against camera attacks. Even though this approach is slightly more complicated than the classical approach, this technique has proven to be accepted by the user in an experimental setting.
E-MAIL SECURITY

Before the middle of the 1970s, cryptography was built entirely on symmetric ciphers. This meant that in order for enciphered communication to take place, a secret key needed to be exchanged beforehand over a secure out-of-band channel. One way of doing that was to send a trusted courier to the party with whom one intended to communicate securely. This procedure addressed two important issues: the secret key exchange and the implicit authentication of the exchanged keys. Once established, the keys could be used to secure communication against passive and active attacks until the key was expected to become or became compromised.

When asymmetric cryptography (Diffie & Hellman, 1976; Rivest, Shamir, & Adleman, 1978) was invented in the 1970s, it tremendously simplified that task of key exchange, and gave birth to the concept of digital signatures. Asymmetric cryptography did not solve the problem of authenticating keys per se. Although we now can exchange keys securely in the clear, how could one be certain that a key actually belonged to the alleged sender? Toward a solution to this problem, Loren Kohnfelder (1978) invented the public key certificate, which is a public key and an identity, signed together in a clever way with the private key of a key introducer whom the communicating parties need to trust. This idea gave rise to the notion of a public key infrastructure (PKI). Some existing models of public key infrastructures are the OpenPGP Web of Trust model (RFC 2440) and the increasingly complex ITU Recommendation X.509-based PKIX model (RFC 3280) (Davis, 1996; Ellison, 1996, 1997; Ellison & Schneier, 2000).

In applications such as electronic mail, building trust in certificates, exchanging keys, and managing keys account for the majority of the interactions and decisions that interfere with the goal-oriented tasks of a user and that the user has difficulty understanding (Davis, 1996; Gutmann, 2003; Whitten & Tygar, 1999). At the same time, the majority of users only has limited understanding of the underlying trust models and concepts (Davis, 1996) and a weak perception of threats (see previous discussion). Consequently, they avoid or improperly operate the security software (Whitten & Tygar, 1999).

In the safe staging approach, security functions may be grouped into stages of increasing complexity. A user may begin at a low stage and progress to a higher stage once the user understands and masters the security functions at the lower stages. The safe-staging approach was proposed by Whitten & Tygar (2003), who also pioneered research on the usability of mail security by analyzing users’ performances when operating PGP (Whitten & Tygar, 1999).

SYSTEM SECURITY

Computer systems progressed from single user systems and multi-user batch processing systems to multi-user time-sharing systems, which brought the requirement to sharing the system resources and at the same time to tightly control the resource allocation as well as the information flow within the system. The principal approach to solving this is to establish a verified supervisor software also called the reference monitor (Anderson, 1972), which controls all security-relevant aspects in the system.

However, the Internet tremendously accelerated the production and distribution of software, some of which may be of dubious origin. Additionally, the increasing amounts of so-called malware that thrives on security flaws and programming errors lead to a situation where the granularity of access control in multi-user resource-sharing systems is no longer sufficient to cope with the imminent threats. Rather than separating user domains, applications themselves increasingly must be separated, even if they run on behalf of the same user. A flaw in a Web browser should
not lead to a potential compromise of other applications and application data such as the user’s e-mail client or word processor. Despite efforts to provide solutions to such problems (Goldberg et al., 1996) as well as the availability of off-the-shelf environments in different flavors of Unix, fine-grained application separation has not yet been included as a standard feature of a COTS operating system.

Even if such separation were available, malicious software may delude the user into believing, for example, that a graphical user interface (GUI) component of the malware belongs to a different trusted application. One means of achieving this is to mimic the visual appearance and responses of the genuine application. One typical example would be a fake login screen or window. Assurance that a certain GUI component actually belongs to a particular application or the operating system component requires a trusted path between the user and the system. For instance, a secure attention key that cannot be intercepted by the malware may switch to a secure login window. While this functionally is available in some COTS operating systems, current GUIs still provide ample opportunity for disguise, a problem that also is eminent on the Web (Felten, Balfanz, Dean, & Wallach, 1997). One approach to solving this problem for GUIs is to appropriately mark windows so that they can be associated with their parent application (Yee, 2002). One instance of a research prototype windowing system designed with such threats in mind is the EROS Trusted Window System (Shapiro, Vanderburgh, Northup, & Chizmadia, 2003).

FUTURE TRENDS

Mobile computing and the emergence of context-aware services progressively are integrating into new and powerful services that hold the promise of making life easier and safer. Contextual data will help the user to configure and select the services the user needs and even might elicit support proactively. Far beyond that, Ambient Intelligence (IST Advisory Group, 2003) is an emerging vision of dynamically communicating and cooperating appliances and devices in order to provide an intelligent surrounding for tomorrow’s citizens. Radio frequency identification transmitters (RFID) already have been discussed with respect to their implications on privacy (Weis, 2004). Certainly, one person’s contextual awareness is another person’s lack of privacy (Hudson & Smith, 1996). In the future, the development of powerful and usable security concepts for applying personal information to the context and vice versa is one of the greatest challenges for today’s security engineers and human-computer interaction research (Ackerman, Darell, & Weitzner, 2001). To accomplish this task seems crucial for future acceptance and for chances of such technologies without them becoming a “privacy Chernobyl” (Agre, 1999).

CONCLUSION

The view of the user as the weakest link and potential security danger finally has turned out to be an obsolescent model. Security engineers and perhaps, more importantly, those people who are responsible for IT security have noticed that working against the user will not do, and instead, they have decided to work with and for the user. During the past years, an increasing number of research has focused on the issue of making security usable, addressing the traditional fields of authentication, communication, and e-mail and system security. This article has given a brief overview of some of the work done so far. In order to make information technology more secure, the user is the central instance. The user must be able to properly use the security mechanisms provided. Therefore, understanding users’ needs
Human-Computer Interaction and Security

and identifying the reasons that technology fails to convince users to employ it is crucial. The first part of this article summarized some work done by Dourish, Weirich, and Sasse that provided important insights. But much work still has to be done.

Future technology will build even more on the integration and sharing of heterogeneous sources of information and services. The tendency toward distributed and location-based information infrastructures will lead to new security problems. Feeling safe is an important aspect of acceptance. The success of tomorrow’s systems also will depend on the user’s feeling safe while sharing information and using services, which has already been shown during the first stage of e-commerce. Therefore, making security usable is an important aspect of making security safe.

REFERENCES


Human-Computer Interaction and Security


KEY TERMS

Asymmetric Cryptography: A data encryption system that uses two separate but related encryption keys. The private key is known only to its owner, while the public key is made available in a key repository or as part of a digital certificate. Asymmetric cryptography is the basis of digital signature systems.

Public Key Infrastructure (PKI): The public infrastructure that administers, distributes, and certifies electronic keys and certificates that are used to authenticate identity and encrypt information. Generally speaking, PKI is a system of digital certificates, certification authorities, and registration authorities that authenticate and verify the validity of the parties involved in electronic transactions.

Shoulder Surfing: The practice of observing persons while entering secret authentication information in order to obtain illegal access to money or services. This often occurs in the context of PIN numbers and banking transactions, where shoulder surfing occurs together with the stealthy duplication of credit or banking cards.

Social Engineering: The technique of exploiting the weakness of users rather than software by convincing users to disclose secrets or passwords by pretending to be authorized staff, network administrator, or the like.

Spoofing: The technique of obtaining or mimicking a fake identity in the network. This can be used for pretending to be a trustworthy Web site and for motivating users (e.g., entering banking information), pretending to be an authorized instance that requests the user’s password, or making users accept information that is believed to come from a trusted instance.

Types of Authentication: Authentication generally can be based on three types of informa-
tion: by some thing the user has (e.g., bank card, key, etc.), by something the user knows (e.g., password, number, etc.), or by something the user is (e.g., biometric methods like fingerprints, face recognition, etc.).

This work was previously published in the Encyclopedia of Human Computer Interaction, edited by C. Ghaoui, pp. 287-294, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 1.26
Security for Ubiquitous Computing

Tobias Straub
Fraunhofer Institute for Secure Information Technology, Germany

Andreas Heinemann
Technical University of Darmstadt, Germany

ABSTRACT

Taking typical ubiquitous computing settings as a starting point, this chapter motivates the need for security. The reader will learn what makes security challenging and what the risks predominant in ubiquitous computing are. The major part of this chapter is dedicated to the description of sample solutions in order to illustrate the wealth of protection mechanisms. A background in IT security is not required as this chapter is self-contained. A brief introduction to the subject is given as well as an overview of cryptographic tools.

INTRODUCTION

Mark Weiser’s vision of ubiquitous computing (UC) raises many new security issues. Consider a situation where a large number of UC peers interact in a spontaneous and autonomous manner, without previously known communication partners and without a common security infrastructure. Such extreme conditions make it difficult to apply established security methods that have been tailored for “classical” information technology (IT). Virgil Gligor emphasizes this by comparing the Internet cliche where “processing is free and physically protected, but communication is not” with the new cliche about UC where “neither processing nor communication is free and physically protected” (Gligor, 2005).

This chapter gives a systematic introduction into the field of security for UC. It is structured as follows: first (the section “Four UC Settings”), we illustrate the diversity of UC systems and applications in order to raise the reader’s awareness of security questions. Our discussion focuses on four representative UC settings. For each of them, a characteristic application scenario is given. The diversity of the settings, for example, concerning presumed knowledge about components or
system complexity, has direct implications for which mechanisms are appropriate to achieve a desired IT security objective.

The section “A Taxonomy of UC Security” starts with an introduction to basic terminology explaining the objectives of IT security as well as the threat model. UC security can be regarded from two different viewpoints: The special characteristics of UC may lead to known and new security risks, if not addressed appropriately. In addition, limitations concerning resources and infrastructure pose a number of challenges in order to achieve desired security objectives.

A compact overview of cryptographic tools suitable to enforce those objectives is given in the section “Overview of Cryptographic Tools”. A brief explanation of cryptographic primitives like ciphers, hash functions, and signatures is provided for the nonspecialist reader. A discussion of the potential and limitations of cryptography in UC concludes the section.

After having learned the particularities and challenges of UC systems with respect to security, we look at sample solutions to mitigate these limitations. A selection of elaborated approaches for secure UC systems is given in the section “Sample Solutions”. They tackle the issues of privacy and availability as well as the establishment of secure communication.

A brief summary and a list of references for further reading conclude this chapter.

FOUR UC SETTINGS

In order to pave the way for the development of a systematic view on UC characteristics and limitations in the section “A Taxonomy of UC Security”, we sketch four representative settings. Each setting exhibits one or more security-related properties. They are termed mobile computing, ad hoc interaction, smart spaces, and real-time enterprises.

Mobile Computing

Mobile computing supports mobile users with connectivity and access to services and backend systems while being on the move. A synonymous term is nomadic computing, emphasizing the goal of providing a working environment more or less equivalent to that of a desktop user. The widespread availability of cellular networks and 802.11 WiFi allows a field worker to connect to an arbitrary service on the Internet or to the company’s backend at almost any place and at any time.

Mobile computing relies on a given infrastructure managed by a provider, for example, a cellular network company. This fact has implications for security: In order to access a service, a user needs to register with a provider. Thus, the user group is closed and the provider controls access to the infrastructure. In addition, users are not able to act in an anonymous manner.

Mobile devices can easily get lost, for example left behind in the proverbial taxi (see http://www.laptopical.com/laptops-lost-in-taxi.html). In case of theft, an attacker might be able to impersonate the legitimate device owner or learn her private data like business contacts or personal email. This physical threat is given whenever mobile devices are considered.

Scenario 1: The Mobile Salesman

While on the road, a salesman needs to regularly download up-to-date client reports from his company’s databases. His laptop is equipped with several wireless communication interfaces which can be used to connect via different service providers depending on what kind of service/infrastructure is available.

At the client’s office, there is a WiFi network the salesman can access. There are also some networked printers available for guests. However, it is unclear to what extent the infrastructure can be trusted.
This scenario raises three major issues: Firstly, the protection of communication (from the salesman's device to the company backend) over potentially insecure channels. Secondly, the secure storage of company internals on his laptop. Secure tunneling, VPNs (virtual private networks), and hard disk encryption are standard technologies in this field. The necessary cryptographic tools are explained in the section “Overview of Cryptographic Tools”. However, the aspect of secure device association, discussed, for example, by Balfanz, Smetters, Stewart, and Wong (2002), is something new. Later in this chapter, we answer the following questions: Is there a way to securely send a confidential document over the air to a printer located in the office? Does it help if the salesman selects a printer close to him equipped with a secondary communication interface?

**Ad Hoc Interaction**

In contrast to mobile computing, the second setting does not rely on an infrastructure provider. Instead of that, UC devices build the infrastructure on their own by establishing temporary, wireless, and ad hoc communication links between them. On the application layer, they expose a spontaneous interaction behavior. A typical characteristic is the lack of a central instance allowing or restricting participation. A priori, there are no given or managed user groups; all devices are free to join. Plus, users and devices might act anonymously.

Here we illustrate a collaboration scenario based on spontaneous interaction. This type of communication is typical for *Opportunistic Networks* which are discussed in depth in the chapter “Opportunistic Networks”.

**Scenario 2: Passive Collaboration in Opportunistic Networks**

*In an Opportunistic Network, passers-by exchange information, for example digital advertisements (Straub & Heinemann, 2004), while being colocated. After an initial configuration, devices interact autonomously and without users’ attention. Information dissemination is controlled by profiles stored on the users’ devices. Such a profile expresses a user’s interest in and knowledge about some pieces of information to share.*

The particularities of ad hoc interaction pose numerous security challenges. On the one hand, devices do not already know each other when they start to communicate. On the other hand, personal data is kept on the devices and exchanged with strangers. As a consequence, privacy is inherently at risk if systems are not designed carefully.

**Smart Spaces**

Smart spaces, which form our third UC setting, emphasize user-friendliness and user empowerment as well as support for human interactions. Interaction within a smart space happens in an unobtrusive way. The use of contextual information (see chapter “Context Models and Context Awareness”) plays also an important role here. Sometimes, it is assumed that users carry some type of digital identification and/or other devices with or on them.

Due to the sensing and tracking capabilities of a smart space, user privacy is at stake. Location privacy is an important field of UC research; an overview is given by Görlach, Heinemann, Terpstra, and Mühlhäuser (2005). In addition, due to the volatile nature of smart spaces, concepts like trust (see Chapter 15) and reputation play an important part in these kinds of applications. We take patient monitoring in a hospital as an example to illustrate a smart space.

**Scenario 3: Patient Monitoring**

*In a hospital, all records of patients are digitally stored and maintained in a central database. Records are updated with the results of physical*
examinations or continuous monitoring. Husemann and Nidd (2005) describe a middleware capable of integrating a wide range of medical analyzers that have a common wireless interface. Consider a battery driven heartbeat monitor which is attached to the body and sends measurements to the database. The data can be used as well for a patient surveillance system that triggers an alarm in case of an anomaly.

This scenario raises a number of security issues: for new patients the heartbeat monitor has to be unambiguously and securely associated with their record. The data from the monitor needs to go to the right record, communication needs to be protected, and the correctness of the data must be assured. In addition, after a patient leaves the hospital, the heartbeat monitor needs to be detached from the digital record for reuse. The resurrecting duckling (Stajano & Anderson, 1999) security policy framework helps in this setting. It is described in the section “Sample Solutions”.

Real-Time Enterprises

Real-time enterprises, which are defined in the preface of this book, are an effort to leverage UC technology and methods within enterprises. A driving force behind these efforts is the goal of having immediate access to comprehensive and up-to-date information about processes and procedures within an enterprise. This allows management to react very flexibly to variances in the market and to increase customer support and satisfaction. For example, an enterprise with detailed real-time information on all production steps, including delivery statuses of subcontractors, can provide a customer with very accurate information on when an order will be delivered.

Scenario 4: RFID-Based Warehouse Management

Radio frequency identification (RFID) offers a variety of opportunities in tracking goods (see e.g., Fleisch & Mattern, 2005). Suppose all goods stocked at a warehouse are tagged with an RFID transponder. With the corresponding readers integrated into storage racks, the process of stocktaking can be completely automated and inventory information is available in real-time.

The use of RFID allows for more efficient stocking, product locating, and product theft protection, to name a few. The deployment of tags and readers within a company gives rise to security questions, for instance, those of industrial espionage. A competitor, equipped with a mobile RFID reader, might be able to derive useful information about a company by reading their product tags. RFID tags should also be looked at from a privacy point of view as they—under certain conditions—may allow the surveillance and tracking of humans.

RFID privacy is currently also a hot topic in the context of machine-readable travel documents. Privacy is discussed with respect to biometric data like photos and fingerprints stored on the chip. Another question is whether ID cards can be used to create movement profiles of their holder. The reader is referred to Knospe and Pohl (2004) for an overview of RFID security. We come back to this issue in the section “Sample Solutions”.

A TAXONOMY OF UC SECURITY

The beginning of this section provides a compact introduction to IT security by explaining the common objectives and the threats to them. We then formulate two views on UC security issues. In section “First View: UC Characteristics and
Associated Risks,” we explain which and how typical characteristics of UC lead to well-known, but also new risks that demand appropriate countermeasures. UC limitations concerning resources and deployed infrastructures are in focus of the second view in the section “Second View: UC Limitations and Associated Challenges”. These limitations give rise to security challenges that require novel methods, some of which are presented in the section “Sample Solutions”.

Basic Terminology and Objectives of IT Security

Security of IT systems is usually discussed in the following way: First, assets (and their respective value) are identified. The notion of assets covers data as well as hardware. With respect to our four scenarios, data that has to be protected, for instance, comprises:

- Confidential documents (Scenario 1),
- An individual’s habits and preferences (Scenario 2),
- Medical information (Scenario 3),
- The stock list at a warehouse (Scenario 4).

When speaking of the protection of data, we mean that particular attributes are preserved. In the information security community the mnemonics CIA or CIAA are often used to refer to the following fundamental protection objectives:

- **Confidentiality (C):** Refers to the aim of keeping pieces of information secret from unauthorized access.
- **Integrity (I):** Is the requirement that data is safe from changes, be it either accidentally or deliberately.
- **Authenticity (A):** Concerns itself with the genuineness of messages or the identity of entities in a networked system.

These objectives can be achieved by cryptography as we will see below. In this respect, the next objective is different as it typically requires noncryptographic efforts as well.

- **Availability (A):** Means the provisioning of a system’s services to its users in a reliable way.

An attacker may try to prevent legitimate use in a so-called denial-of-service (DoS) attack. Demanding a large portion of (computational and/or network) resources is a strategy to slow down the system. Another flavor of DoS is the blocking of communication links between legitimate parties. Designing a UC system for redundancy and diversity is a standard way towards achieving availability (see e.g., Vogt, 2005). This can be done, for instance, by using peer-to-peer networks, the replication of information, or distributed computing facilities. Intrusion-detection systems, which are a common protection technology for computer systems in an organization, are starting to attract attention in the UC community (Robinson, 2005). Bahl, Chandra, Padhye, Ravindranath, Singh, Wolman (2006) describe a system to fight DoS attacks on corporate wireless networks and to detect malicious mobile devices.

Threat Modeling

Having discussed basic terminology and objectives, we now turn to a common abstract network model to describe security threats: Two parties exchange messages over a channel to which an attacker has access, too. This notion captures any kind of attackers (computer systems, individuals, organizations) and is independent of the data transport medium itself. There is no need to differentiate between the transmission and the storage of data as the latter can be seen as a special case of the model.
We follow the security community's convention in using the names Alice and Bob for the legitimate actors (instead of simply numbering them serially) and in calling the attacker Mallory (“malicious”). Mallory may change data Alice sends to Bob, may generate her own messages under the name of another person, or simply eavesdrop on their connection. An attack of the latter kind is called passive while the other two are called active. A passive attacker can compromise confidentiality at best, but an active one—who is also called man-in-the-middle (MITM)—targets at all CIAA goals. Acting as a MITM, Mallory sits in between the communication link, making Alice believe she is Bob and spoofs Bob into believing she is Alice.

Attacks are typically directed toward more than one of the before-mentioned security objectives. For instance, Mallory may launch a DoS attack in order to paralyze the system’s defense mechanisms.

The risk that a system may become compromised is proportional to both its vulnerabilities and the threats acting upon it. Risks have to be identified and rated in the light of the corresponding assets’ value. Threat perception can be very subjective: While one individual cares about data emitted by her UC device, which might be linked back to her, another does not. Furthermore, not all objectives are equally important in practice. For instance, a bank primarily has a vital interest in keeping its account data’s integrity, but a research lab emphasizes confidentiality. The choice of algorithms and security also takes into consideration the (assumed) attacker’s strategy and resources, especially computational power.

First View: UC Characteristics and Associated Risks

In order to put the UC vision to work, most scenarios, including the ones described in the last section, rely on wireless communication between nodes. Communication might happen in an ad hoc manner as illustrated in the ad hoc interaction setting; other applications might ask for a wireless multi-hop communication (see sensor networks in the chapter “Wireless and Mobile Communication”). Wireless communication makes eavesdropping very easy as radio signal are usually emitted in all directions. They can be received by anyone in the senders’ vicinity without her noticing it. MITM attacks are feasible in the case of multi-hop wireless communication and they do not even require the attacker to be physically close to the victim. In addition, wireless ad hoc communication bears the risk of impersonation, that is, an attacker might be able to steal a peer’s credential by eavesdropping and use it to access a certain service.

The pervasive nature of UC introduces even more risks. Sensor nodes or RFID tags for example, are physically exposed, unmanaged, and unsupervised. This bears the risks of device and/or data theft as well as device manipulation. As a consequence, access to the data stored on the device must be carefully protected in order to prevent identity theft (Eckert, 2005). The Blackberry PDA for instance is a centrally manageable system that supports a remote “kill” command to erase data on a stolen device (Greiner, 2006).

UC devices are often battery-powered, which allows for a DoS attack called sleep deprivation torture (Stajano & Anderson, 1999): By constantly sending requests to a device, an attacker can quickly drain its battery, thus rendering the device useless. Last but not least, the UC settings offer the capability of tracing objects or humans. This feature is useful in many UC applications like the real-time enterprise setting, but may violate user’s privacy. Networked sensors, for instance, may gather a good deal of personal information that can be used to build user profiles.

The characteristics in UC and their corresponding risks are summarized in Table 1. In the section “Overview of Cryptographic Tools” we cover appropriate countermeasures.
Second View: UC Limitations and Associated Challenges

Our second view on UC security concerns resource and infrastructure limitations of UC and their corresponding challenges. The functionality provided by UC devices varies widely and is limited by a number of parameters. These include device capabilities—concerning, for example, memory capacity, energy supply, CPU power, and the user interface—or connectivity which is characterized among others by network coverage, bandwidth, environmental conditions (like shielding or speed in case of moving devices).

Since a UC device can be less powerful than a desktop computer, it is important to use appropriate, that is, lightweight, security mechanisms. Otherwise, wasting memory and energy resources would achieve the exact opposite effect of keeping the system safe. Observe that a slow CPU might take significant time to carry out sophisticated cryptographic operations which may negatively affect the user experience. User interface capabilities may also limit the choice of appropriate methods, too: For example, a mobile device with a voice only user interface such as described by Ait-enbichler, Kangasharju, and Mühlhäuser (2004), demands different authentication methods to a device that comes with a keyboard. At the bottom end, very small and feeblish UC devices like passive RFID chips are found. These devices do not even come with a user interface at all. A limited user interface makes it difficult to establish what is called a trusted path. This is a mechanism to ensure that users can assert that they are interacting with a genuine system or program instead of one controlled by an attacker (for instance when asked for a password).

The absence of a centralized authority is typical for UC systems like the ad hoc interaction setting presented in section “Ad Hoc Interaction”. This makes it hard to build an infrastructure where UC peers can authenticate each other in order to communicate securely. Another consequence is the lack of common support for trust and policy decision. The following chapter discusses appropriate countermeasures for this issue. Alternative methods for entity authentication based on out-of-band communication schemes are the subject of section “Out-of-Band Channels”.

The limitations corresponding to UC security and their induced challenges are summarized in Table 2. Selected methods that address these challenges are discussed in section “Sample Solutions”.

<table>
<thead>
<tr>
<th>characteristics</th>
<th>risks</th>
</tr>
</thead>
<tbody>
<tr>
<td>communication</td>
<td></td>
</tr>
<tr>
<td>wireless</td>
<td>eavesdropping</td>
</tr>
<tr>
<td>ad hoc</td>
<td>impersonation</td>
</tr>
<tr>
<td>multi-hop</td>
<td>man-in-the-middle attacks</td>
</tr>
<tr>
<td>pervasive nature</td>
<td></td>
</tr>
<tr>
<td>physical exposure</td>
<td>device/data theft, manipulation</td>
</tr>
<tr>
<td>battery-powered</td>
<td>sleep deprivation torture</td>
</tr>
<tr>
<td>traceability</td>
<td>privacy violation</td>
</tr>
</tbody>
</table>

Table 1. First view: Characteristics and risks
Security for Ubiquitous Computing

Table 2. Second View: Limitations and Challenges

<table>
<thead>
<tr>
<th>Limitations</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Infrastructure</strong></td>
<td></td>
</tr>
<tr>
<td>lack of centralized authority</td>
<td>entity authentication, policy</td>
</tr>
<tr>
<td></td>
<td>decision</td>
</tr>
<tr>
<td><strong>Resource</strong></td>
<td></td>
</tr>
<tr>
<td>limited CPU power, few/no</td>
<td>algorithm implementation,</td>
</tr>
<tr>
<td>memory, limited power supply</td>
<td>protocol design</td>
</tr>
<tr>
<td>user interface limitations</td>
<td>trusted path</td>
</tr>
</tbody>
</table>

OVERVIEW OF CRYPTOGRAPHIC TOOLS

This section provides a compact overview of cryptographic tools which are suitable to enforce the security objectives introduced in the section “A Taxonomy of UC Security”. It serves as a basis for the following discussion of the solutions addressing UC characteristics and challenges. Cryptographic primitives, namely symmetric and public key cryptosystems, hash functions, and authentication schemes, are the building blocks of more complex techniques and protocols. We also introduce some general rules of cryptography and reason about security parameters. At the end of this section we turn to the potentials and limitations of cryptography in a UC setting. Readers with previous knowledge in the area of cryptography may directly proceed to section “Potential and Limitations of Cryptography in UC.”

Symmetric Cryptosystems

Messages can be encrypted to prevent disclosure of their content and ensure confidentiality. Encryption is a transformation that renders plaintext, that is, the original message in its readable form, into ciphertext which is unintelligible to an outsider. The reverse transformation is called decryption. In order to give Alice and Bob a competitive edge, there must be some information Mallory does not have. The encryption and decryption function of all modern cryptosystems (or ciphers) are parameterized by keys. A key is a short fixed-length piece of data (while the plaintext may be of arbitrary length). It has to be kept secret from Mallory as only the correct key allows successful decryption of the ciphertext.

As early as the 19th century, cryptographer Kerckhoff claimed the following important law: A cryptosystem’s strength should not be based on the assumption that its algorithm is kept secret, but only on the attacker’s uncertainty regarding the key. An elementary security requirement is the fact that it should be unfeasible for Mallory to simply try out all possible keys by a brute-force attack. This imposes a lower bound on the cardinality of the set of all possible keys and demands that keys be generated randomly.

The distinction whether the same key is used for encryption and decryption or not may seem marginal at first sight, but is in fact crucial. Symmetric cryptosystems require two or more communication partners to know the same shared key. Modern symmetric cryptosystems use keys of bitlength 112, 168 (3DES) or 128, 192, 256 (AES). As a rule of thumb, keys shorter than 80 bits should be used with great care. A good place to look up reliable key lengths based on cryptographic analysis is the Website http://www.keylength.com.
On this occasion, we exhibit a problem common to all symmetric ciphers: The use of shared keys implies a secure key distribution step before the actual communication takes place. Key distribution is often a problem as it needs out-of-band mechanisms, that is, an additional communication channel not accessible to the attacker. We will come back to this issue in the context of UC in the section “Out-of-Band Channels”. Another downside of shared keys is that the risk grows proportional to the number of group members. If one device falls into Mallory’s hands, she will be able to read all messages of the group. A real-world example for the use of shared keys is 802.11 WEP (“Wired Equivalent Privacy”, see, for example, Eckert, 2006).

**Asymmetric Cryptosystems and PKI**

The key distribution problem can be avoided by using asymmetric cryptosystems, which are synonymously called public key cryptosystems. In this case, encryption and decryption use different keys, such that these functions can be effectively separated. The respective keys are denoted public key and private key. Such a key pair has to satisfy certain mathematical conditions depending on the particular cryptosystem. Bob solely has to keep his private key secret, but can make the corresponding public key available to Alice or everybody else in the world—including Mallory—as it is infeasible to compute the private key from the public key. Public key encryption is thus an example of a so-called one-way function since the ciphertext can be computed efficiently, while inversion is infeasible. More specifically, public key encryption falls into the category of one-way functions with trapdoor as knowledge of the private key allows the inversion of the mapping. We will see below that asymmetry is an important pillar of modern cryptography.

The RSA cryptosystem, named after R. Rivest, A. Shamir, and L. Adleman (1978), is the most prominent example of an asymmetric cryptosystem. Typical key lengths are 1024, 2048, or 4096 bit. Its security is based on the presumed intractability of the factorization of large composite numbers. The ElGamal cryptosystem in contrast is based on the problem of computing discrete logarithms in finite cyclic groups (a typical group size is $2^{160}$ or $2^{230}$). A standard construction for ElGamal is to use a subgroup of the multiplicative group of $\mathbb{Z}/p\mathbb{Z}$, where $p$ is a prime (typically 1024 or 2048 bits in length). Elliptic curves (EC) over finite fields can also serve as a mathematical structure for ElGamal encryption (see, e.g., Hankerson, Menezes, and Vanstone, 2004) for an introduction to EC cryptography.

Observe that public key cryptography alleviates the problem of confidential secret key distribution for the price of authentic public key distribution in the following way: Alice who wants to send a private message to Bob needs to have his genuine public key $K_B$. If Mallory manages to foist her key $K_M$ on Alice, the message would be encrypted for her and not for Bob. Digital certificates are a common solution to this problem. A certificate is a statement linking a public key to an identity (e.g., a user’s name, a device address, or a pseudonym). In order to prevent certificates from being forged they are digitally signed (see the following section for an explanation of digital signatures). Certificates form the basis of a so-called public key infrastructure (PKI), an environment where the authenticity of public keys is assured.

**Hash Functions**

In order to ensure data integrity, some kind of redundancy has to be added to the payload. A modification detection code (MDC) is a hash algorithm, that is, a function that compresses bitstrings of arbitrary finite length to bitstrings of fixed length. State-of-the-art algorithms like RIPEMD-160 (Dobbertin, Bosselaers, & Preneel, 1996) or SHA-1 (Eastlake & Jones, 2001) produce outputs with a length of 160 bit, which should be
considered the lower bound due to security reasons. To be useful for integrity protection, a hash function has to be 2nd preimage resistant: Given an input \( x \) that hashes to \( h(x) \), an attacker must not be able to find a value \( y \neq x \) such that \( h(y) = h(x) \). Such a pair of different inputs that result in the same hash value is called a collision. Collisions always exist due to the pigeon-hole principle since the co-domain is finite while the domain is infinite. Cryptographic hash functions must be collision-resistant, that is, finding concrete collisions must be computationally infeasible. This allows us to protect data of arbitrary length against modifications by storing \( h(x) \), a small piece of information, in a safe place. A widespread application of this mechanism is integrity-protected software distribution via download mirror sites on the Internet. Flipping a single bit in \( x \) results in a hash value that differs in about half of the bits from \( h(x) \) if \( h \) is one of the common hash functions.

Message authentication codes (MACs) are hash functions that are additionally parameterized with a secret key. Assume that Alice and Bob share a key \( k \), which for instance has been established in a key exchange scheme as explained below. Alice adds a MAC \( h_k(x) \) to her message \( x \) to Bob. From the pair \( (x, h_k(x)) \) Bob can tell that the message came from her since it is infeasible for Mallory to create a pair \( (y, h_k(y)) \) without knowing \( k \). Remember that a MDC does not have this property, since Mallory may change \( x \) to \( y \) and compute the (unkeyed) hash function \( h(y) \). MACs in turn not only provide data origin authentication, but also integrity as Bob could detect a modification \( x' \neq x \) because the received \( h_k(x) \) would not match the value \( h_k(x') \) he computed himself. Each MDC \( h \) can be extended to a MAC in the following way: On input \( x \), compute \( h(k \parallel p_1 \parallel h(k \parallel p_2 \parallel x)) \) where \( k \) is the key, \( p_1, p_2 \) are constant padding strings, and \( \parallel \) denotes concatenation. This generic construction is called HMAC (hash-based MAC), see Krawczyk, Bellare, and Canetti, 1997 for implementation and security details.

Digital Signatures

MACs are useful for guaranteeing the authenticity of a communication link. However, they do not provide a transferable proof of authorship: As the parties at both ends of the channel share knowledge of the key, a “MACed” message \( (x, h_k(x)) \) could originate from either of them. As a matter of fact, a third party cannot deduce its author. Transferable proofs of authorship are required to model digital workflows, for example orders in electronic commerce that must be verifiable by multiple parties (possibly including a judge in order to arbitrate in a dispute). Digital signature schemes are a means to provide such evidence. They are also implemented with public key cryptography. Signature verification is a one-way function with trapdoor, referring to the ability to create signatures.

The RSA signature scheme is based on the same mathematics as the RSA cryptosystem. Signing corresponds to decryption while verification corresponds to encryption. Digital Signature Algorithm (DSA) and its EC-based variant EC-DSA are alternative schemes which are not yet as prevalent as RSA. EC-DSA is more efficiently computable than RSA and uses shorter keys for comparable strength.

Signatures are, in fact, computed over the hash value \( h(x) \), not the message \( x \) itself. One reason is the reduction of computational costs associated with public key cryptography. Secondly, the one-way property of the hash function prevents so-called existential forgery of signatures (see e.g., Buchmann, 2004). However, \( h \) must be collision-resistant as in the case where \( h(x) = h(x') \), a signature of \( x \) is always a signature of \( x' \), too.

Messages that comprise declarations of intent should be legally binding for the originator. On the one hand, this implies that the originator commits himself to the content in a way that he cannot deny his intent later. On the other hand, no one other than the legitimate person should be able to
make a commitment in his name. This property is called non-repudiation. Non-repudiation can be achieved with the help of dedicated hardware, for example a smartcard or TPM (Trusted Platform Module, see e.g., Eckert, 2006) chip that keeps the signing key under the sole control of its owner.

Observe that deniability, the opposite of non-repudiation, may indeed be a desirable security goal sometimes, for instance in order to cope with censorship in UC networks. In such a setting, MACing should be preferred to signing of messages.

Hybrid Encryption and Key Exchange

In practice, asymmetric and symmetric cryptosystems are combined in hybrid encryption in the following way: Symmetric cryptosystems typically have a much higher throughput compared to public key systems with a comparable security level. Therefore Alice first encrypts the plaintext with a symmetric cryptosystem using a random, ephemeral session or transaction key. Then she transmits the symmetric key to Bob in a confidential way by encrypting it with his public key (if the message has multiple recipients, they all obtain the same ciphertext accompanied by the same symmetric key encrypted with their individual public key). The mechanism of distributing key material is called key exchange. The term key agreement is often used synonymously, but may sometimes emphasize the fact that both parties contribute to the generation of the key. The Diffie-Hellman (DH) protocol is the most famous interactive key exchange scheme. It allows two parties linked by an authentic, but not necessarily private, channel to agree on a common secret—which seems counter-intuitive at first sight. Note that unauthenticated DH is inherently prone to MITM attacks. Authenticated DH can be implemented by having each party digitally sign its protocol messages. The mathematics and security properties of DH are closely related to the ElGamal cryptosystem.

An authentic key agreement protocol like the DH scheme allows parties to achieve forward secrecy, that is, to limit the possible impact of a broken key: If a long-term signing key falls into Mallory’s hand, this does not leak the plaintext of past communications Mallory may have recorded. On the other hand, a broken ephemeral symmetric key only affects the particular messages encrypted with it. Forward secrecy allows the parties to use relatively short symmetric keys provided they are changed frequently. When this trade-off is made in a UC scenario, the cost reduction for symmetric encryption due to smaller key sizes has to be balanced against the overhead for key exchange.

Potential and Limitations of Cryptography in UC

Having outlined the general tools of cryptography, we conclude this section by a discussion of computational costs. Since energy consumption is a serious issue in UC, we direct the reader’s attention to Table 3. These figures are the outcomes of experiments with a 206 MHz Compaq iPAQ H3670 described by Potlapally, Ravi, Raghunathan, and Jha (2003). The table shows the setup and current costs for the families of algorithms we explained above. Observe that the costs of hashing and symmetric encryption are given in μJ, whereas mJ is the unit for digital signatures.

Due to its output size of only 128-bit, we included the hash function MD5 for comparison although its use is no longer recommended. The figures for the AES algorithm refer to the standard operation mode CBC (see e.g., Buchmann, 2004). AES is a so-called iterative block cipher. Its computation is organized in multiple rounds implementing the same subroutines, but with different round keys. The round keys are derived from the actual (128, 192, or 256 bit) key. This process is called key scheduling.
The table shows the signature algorithms RSA, DSA, and EC-DSA with their respective security parameters. These three combinations are considered of having comparable strength. While the costs for the respective setup do not differ too much, the costs for signing and signature verification vary significantly. Obviously, EC-DSA is a better choice than DSA as each operation is cheaper. However, the choice between RSA and EC-DSA is determined by the expected number of operations of the UC system. If the number of signature verifications compared to signature generations dominates, the RSA signature scheme is clearly favorable compared to EC-DSA.

However, concerning key exchange, EC-DH definitely outperforms standard DH (where computations are done in $\mathbb{Z}/\mathbb{Z}_p$) with the same security level. In order to achieve roughly the same costs as EC-DH, DH keys must be restricted to 512 bits in length.

We have not mentioned asymmetric encryption so far. Since the same mathematical formula is used for the RSA cryptosystem and the RSA signature scheme, the corresponding numbers can be looked up in the table. This shows that in some UC situations with constrained devices, asymmetric cryptosystems may quickly become impractical while symmetric encryption is cheaper by magnitudes. Contemporary public key algorithms are not very well suited for the capabilities and needs of UC (Buchmann, May, & Vollmer, 2006). A back-of-the-envelope calcula-
tion shows that a pocket PC’s battery with a 1500 mAh capacity and a 5V voltage would have lost 20% of its charge after 5000 executions of a DH protocol or 10000 RSA signatures.

These facts illustrate that it is worthwhile to precisely analyze the security demands of a UC system since choices on the cryptographic layer directly affect energy consumption. A comprehensive examination of energy costs is provided by Seys (2006). Lightweight UC cryptography has also been the subject of recent research activities, for example, by the ECRYPT Network of Excellence in Cryptology (http://www.ecrypt.eu.org). This issue will surely gain even more momentum with the deployment of RFID tags for industrial applications as we already sketched in Scenario 4.

Lightweight cryptography has several aspects. On the one hand, it comprises the design choice of algorithms and efficient engineering to implement them in hard- and software. On the other hand, it requires a re-assessment of the threat model by assuming a less powerful opponent and ruling out some particular attacks which are considered less likely. This may imply that weaker security mechanisms, like weak authentication (discussed next), or lower security parameters are acceptable for low-value transactions (cf. the discussion of forward secrecy).

Cryptography offers a wealth of ideas that can be applied to UC: For instance, traditional protocols can be enhanced to provide privacy protection using mechanisms like the private authentication scheme explained below. We also point out that threshold cryptography (see for example, Gemmell (1997) for an introduction) is a valuable tool for UC that merits further study. For instance, it can be used in sensor networks that are prone to so-called Byzantine faults, that is, there is a risk that devices fall into the attacker’s hands and then act incorrectly. Provided that the number of such “Byzantine” nodes is lower than an adjustable bound, the attacker learns nothing about a secret (e.g., a common key for signing responses) shared among the nodes. Interestingly, this statement holds in an unconditional sense, that is, for an adversary with unlimited computational power.

**SAMPLE SOLUTIONS**

In this section five example solutions for secure UC systems are described. References to the literature and UC applications as well as links to our scenarios are given in the text.

At first, we present two strategies of privacy-enhancing technologies: the technical concept of anonymity and a policy-driven approach. While anonymity is favorable from a privacy perspective, it nevertheless leads to new threats as malicious parties may benefit from it, too. One such threat is the so-called Sybil attack which is discussed in the section “Fighting Sybil and DoS Attacks”. We explain the technique of proof-of-work as a countermeasure which can also be used as a defense against DoS. A typical situation in UC is the lack of a priori trust relations among the participants. In order to cope with such a setting, a secure communication channel has to be bootstrapped somehow. The section “Bootstrapping Secure Communication” treats this aspect from a theoretical point of view. Above all, the resurrecting duckling security policy model is explained here. The following section gives several real-world examples for out-of-band channels used to set up secure UC communication. Finally we touch on the hot topic of RFID security and discuss in particular privacy concerns and protection mechanisms of electronic travel documents.

We deliberately omitted reputation systems from the set of examples, as this topic is the subject of the following chapter. Trusted Computing is also an interesting direction of current UC research, but would go beyond the scope of this chapter. We refer the interested reader to the work of Hohl and Zugenmaier (2005).
Privacy-Enhancing Technologies

A large number of UC environments rely on sensing and tracking technologies of users and devices in order to carry out their tasks. For example, in order to provide location-based services, a user’s position has to be determined beforehand. But even without a given UC service in place, the fact that most communication takes place over a wireless link opens the door for attacks based on traffic analysis. All communication patterns that can be successfully linked to a human jeopardize a user’s privacy. In this sense, the notion of confidentiality has to be extended to message source confidentiality and/or message destination confidentiality. This property of hiding the fact that particular communication relationships exist at all is also called sender or recipient anonymity, respectively. Anonymity is significantly more difficult to achieve than message content confidentiality, as it requires careful design of communication protocols. Examples in the Internet setting include mix networks (Chaum, 1981), onion routing (see http://tor.eff.org for a popular implementation), or anonymous re-mailers. Those protocols however cannot be directly adapted to the UC world, since they typically make assumptions about connectivity and resources. A first step in this direction is the idea of “Mist Routing” (Al-Muhtadi, Campbell, Kapadia, Mickunas, & Yi, 2002). “Mix Zones” are a variant of the concept of mix networks to locations (see Beresford & Stajano, 2004).

As mentioned before, anonymity is of particular interest in UC when it comes to achieving location privacy, that is, to prevent people bearing devices emitting radio signals from being tracked without noticing it. We discuss two different strategies to preserve a user’s privacy. First, a number of technical means are presented in order to blur data that could later be used to identify a person/subject. Second, the policy-based approach to achieve user privacy as proposed by Langheinrich (2002).

Blurring Data

Gruteser and Grunwald (2003) propose a middleware-based approach to blur location information from clients before passing the location information onwards to a location-based service provider. They assume that clients communicate their position as very precise location information to a location server. Position is determined on the client itself, for example, via GPS or by the wireless service provider through signal triangulation. Location-based service providers access location information through the location server (see Figure 1).

A client specifies that he wants to be indistinguishable from at least \( k - 1 \) other clients within a given area and time frame. In other words, the clients want to stay \( k \)-anonymous. To reach \( k \)-anonymity, their algorithm, an adaptive form of a quadtree-based algorithm, adjusts the resolution of location information along spatial and temporal dimensions in order to meet the specified anonymity constraints, say \( k \).

Providing anonymity requires careful system design as identification may happen on each network layer (Avoine & Oechslin, 2005). As a

Figure 1. Anonymizing middleware for location-based services
practical consequence, user-controlled pseudonyms on the application layer are pretty useless if the devices itself can be re-identified. This is the case when static IP or Media Access Control addresses are used. A proposal for temporary addresses for anonymity on the data link layer is made in Orava, Haverinen, and Honkanen (2002). UC environments like scenario 2 are particularly suitable for addresses that are picked at random, since the number of devices that are connected to each other at the same time is by magnitudes smaller than in an Internet setting. This guarantees, with a high probability, that device addresses do not coincide.

Private Authentication

The following example illustrates how cryptographic protocols can be modified to provide location privacy on the application layer. In typical schemes for mutual authentication like the one used in TLS/SSL (Dierks & Rescorla, 2006), the parties reveal their identity during the handshake. As a consequence, Mallory may pretend to be a legitimate party and start a handshake in order to see who is in her vicinity. She may also eavesdrop while Alice and Bob are authenticating each other and learn about their communication relation. Private authentication tackles this problem. The approach of Abadi (2002) allows Alice to prove her identity to some self-chosen set \( S \) of peers in order to establish a private and mutually authenticated channel. Entities outside of \( S \) cannot detect Alice’s presence while entities inside \( S \) do not learn more than their own membership. Without loss of generalization, we restrict ourselves to the case where Bob is the only member of \( S \). If the set \( S \) contains more than one element, Alice simply goes through parallel executions of the protocol for each element. In order to start a communication with Bob, Alice broadcasts the plaintext “hello” accompanied with

\[
 c := \text{Enc}_{K_B}(\text{“hello” ||} K_A || \text{Sig}_{K_A^{-1}}(K_A || K_B || K || t) ).
\]

Here \( K_B \) is Bob’s public key used for encryption, \((K_A,K_A^{-1})\) denotes her own key pair, \( K \) a session key, and \( t \) is a timestamp. As Bob is the only entity who knows the private key \( K_B^{-1} \), he is able to extract from \( c \) the sender’s identity (which is unforgeable due to the digital signature). Assuming that Alice is on his whitelist of peers, he answers with a message protected by \( K \). Bob’s presence in turn cannot be detected provided that the cryptosystem is which-key concealing (i.e., \( K_B \) cannot be deduced from \( c \)) and Bob ensures \( t \)’s recency. Otherwise Mallory could mount a replay attack by sending \( c \) and checking whether there is a corresponding answer. Abadi (2002) also describes a second protocol without timestamps at the price of additional communication rounds. This variant might be useful in cases where one cannot assume synchronized clocks.

A Policy-Based Mechanism

Langheinrich (2002) proposes pawS, a system that provides users with a privacy enabling technology. This approach is based on the Platform for Privacy Preferences Project (P3P, see http://www.w3.org/TR/P3P/), a framework which enables the encoding of privacy policies into machine-readable XML. Making use of a trusted device, the so-called privacy assistant (PA), the user negotiates his privacy preferences with the UC environment. For this purpose, the PA is able to detect a privacy beacon upon entering a UC environment, for example, a smart space. The privacy beacon announces the available services, for example, a printer, or a video camera, with a reference to their data collection capabilities and policies. The PA in turn contacts the user’s personal privacy proxy located on the Internet, which contacts the corresponding service proxies.
Security for Ubiquitous Computing

at their advertised addresses (through the privacy beacon) and asks for the privacy policies of the services located in the smart space. These privacy policies are compared to the user’s privacy preferences and a result may be to decline usage of a tracking service, which will result in disabling a video camera. Further, the work of Langheinrich (2001) includes four design principles for UC environments, the main ideas of which we sum up as follows:

- **Notice**: Since future UC environments will be ideally suited for unnoticed operation, monitoring and tracking, the author proposes some kind of announcement mechanism that allows users to notice the data collection capabilities in their environment. This announcement mechanism should be standardized and agreed on.

- **Choice and consent**: To preserve privacy, a future UC system should offer users the choice of allowing or denying any kind of data collection. A system should respect this choice and operate only on user consent. Since a user will not always explicitly express her will, her choice and the derived consent should be expressed in machine-readable policies.

- **Proximity and locality**: Locality information for collected data should be used by the system to enforce access restriction. This restriction may be based on the location of a user who wants to use the data, for example, a user is able to record an audio stream at a meeting while attending, but not from the opposite side of the world (proximity). Also the audio stream should not be disseminated through the network (locality). This idea is also present in location-limited channels and location-based authentication (see below).

- **Access and recourse**: A system should give users easy access to collected personal information (e.g., using a standardized interface). Also, they should be informed about any usage of their data. Thus, abuse would be noticed.

Following these principles, Langheinrich’s goal is to allow people and systems which want to respect a user’s privacy to behave in such a way. This should help in building a lasting relationship based on mutual trust and respect. However his goal is not to try to provide perfect protection for personal information that is hardly achievable anyway.

**Fighting Sybil and DoS Attacks**

As mentioned in the section “First View: UC Characteristics and Associated Risks”, DoS may be a serious attack in UC. In a world of widely-deployed pervasive computing devices, a phenomenon comparable to unsolicited email as we experience it today in the Internet setting would be disastrous. In this section we describe a mechanism that can be use as a remedy for DoS attacks and for so-called Sybil attacks as well (see Levine, Shields, and Margolin (2006) for an overview). The latter are an inherent threat to recommender or reputation systems, which are the subject of the following chapter. The term Sybil stems from the name of a schizophrenic literary character that has a total of 16 personalities. Analogously, an attacker takes part in a recommender system under multiple identities in order to manipulate the trustworthiness rating of nodes. DoS and Sybil attacks have some similar characteristics which is why they are jointly dealt with in this section.

In a setting where UC nodes can be identified in a reliable way, the risks for both attacks are quite low. Clearly, a necessary condition for a Sybil attack is a node’s ability to remain anonymous. In the DoS case, illegitimate messages can be filtered out by so-called whitelists and blacklists containing accepted and non-accepted entities, respectively. Filtering in its turn can be done on each network layer provided the sender...
In the following example we sketch a mechanism that is similar to the one applied by Hashcash (Back, 1997). Here, the challenge is time-dependent. The proof of work consists of partially inverting a one-way hash function \( h \) by brute-force search. The sender of a message has to find a solution \( X \) to the following equation:

\[
\text{hash} (\text{id} \ || \ t \ || \ X) = 0 \ldots 0 \quad \text{\( \ell \)-times}
\]

Here, \( \text{id} \) is the recipient’s identifier (e.g., a pseudonym), \( t \) is a time stamp and \( h \) is the output of \( h \) shortened to the leading \( \ell \) bits. The tuple \((\text{id}, t, X)\) is sent along with the payload. The following checks are made at the receiver’s side: At first, she checks whether \( \text{id} \) is her own ID and whether the message is suitably recent, that is, \( t \) differs less than a predefined clock skew bound of \( s \) seconds from her local time. Furthermore the above condition must hold. In order to prevent a PoW from being reused, \((t, X)\) can be checked for freshness against a database that caches previous values for \( s \) seconds.

The sender has to try out \( 2^\ell \) possibilities for \( X \) on average, until a solution is found. This implies that the sender has to invoke the hash function \( 2^\ell \) as many times as the recipient. The complexity for the sender grows exponentially in \( \ell \). Depending on the performance of the devices in use, \( \ell \) should be adjusted in a range of, say, 10–20 in practice.

We will revert to the concept of PoW in the following chapter in the context of micropayment schemes.

**Bootstrapping Secure Communication**

Consider the situation where two UC devices want to establish a secure communication channel between each other in the presence of an adversary. Here secure may have the meaning of confidentiality and/or authenticity. As confidentiality is
Achieved by encryption, the parties have to secretly agree on a shared key or exchange their public encryption keys in an authentic way—depending on whether they use a symmetric or asymmetric cryptosystem. The requirements for authenticity are quite similar: a MAC key has to remain secret as well and the public key for the verification of a digital signature has to be authentic.

The bootstrapping of a secure channel thus boils down to either initially obtaining some information in a confidential or authentic way. Observe that, by using public key cryptography, an authentic communication relation can be enhanced to an encrypted one. If there is no (common) central authority like a certification authority or a key distribution center the devices have a trust relation with, they have to resort to direct exchange of cryptographic information via so-called out-of-band communication. This is a well-known concept which is applied, for instance, in the context of secret key distribution by messengers or the verification of public key fingerprints on the phone.

Out-of-band communication requires a secondary link, typically with some technical limitations, but with the huge advantage of being safe from attacks. Such limitations, that make this secondary channel unattractive for general data transfer, include bandwidth, signal strength, error rate, latency, and so forth. What is typically considered a downside simultaneously reduces the risk of attacks on this channel. Balfanz, Smetters, Stewart, and Wong (2002) use the term of location-limited channels and emphasize that humans can be in control of the communication links. These authors assume an authentic out-of-band communication channel while the resurrecting duckling security policy model by Stajano and Anderson (1999) requires a confidential one. Stajano and Anderson use a confidential channel in order to securely pair mobile devices for a certain period of time; they use the term secure transient association. We will discuss their model in greater detail now.

Secure transient association addresses the problem of device authentication in the absence of a central and always available authority, for example a public key infrastructure. Devices authenticate each other and agree upon a shared key by physical contact. For a user, physical contact has the advantage that it is simple to understand and it is clear which devices are involved in the paring. The two devices take different roles:

- A slave (or duckling) obeys a master
- A master (or mother duck) controls a slave

Initially the slave is in state imprintable. As soon as a master has sent a key to the slave—this process is called imprinting—via a physical contact, the slave follows commands only given by the master who owns the key. The slave is now in the state imprinted.

The association can be broken in three ways. First, the master sends a special kill command to the slave. The slave will reset itself into the initial state imprintable. Second, after a predefined time interval has gone by, the slave kills itself. And third, after the completion of a certain transaction, the slave will reset itself to imprintable.

The two states of a slave and its transition are depicted in Figure 2. Note that the slave needs to be constructed in a tamper resistant manner. This guarantees that it is uneconomical for an attacker to attack by artificially causing a slave to die, that is, to force a slave into state imprintable. Otherwise, this would allow an attacker to imprint the slave using its own key and gain control over the slave. We quote Stajano’s four formal principles that make up the resurrecting duckling security policy (Stajano, 2002):

1. **Two state principle:** The entity that the policy protects, called the duckling, can be in one of two states: imprintable or imprinted. In the imprintable state, anyone can take it over. In the imprinted state, it only obeys its mother duck.
2. **Imprinting principle:** The transition from imprintable to imprinted, known as imprinting, happens when a principal, from then on known as the mother duck, sends an imprinting key to the duckling. This must be done using a channel whose confidentiality and integrity are adequately protected. As part of the transaction, the mother duck must also create an appropriate backup of the imprinting key.

3. **Death principle:** The transition from imprinted to imprintable is known as death. It may occur under a very specific circumstance, defined by the particular variant of the resurrecting duckling policy model that one has chosen. Allowable circumstances, each corresponding to a different variant of the policy, include:
   - Death by order of the mother duck.
   - Death by old age after a predefined time interval.
   - Death on completion of a specific transaction.

4. **Assassination principle:** The duckling must be constructed in such a way that it will be uneconomical for an attacker to assassinate it, that is, to cause the duckling’s death artificially in circumstances other than the one prescribed by the Death principle of the policy.

An appropriate key backup of the imprinting key is necessary in order to keep control of the slave in case the master is broken. Coming back to the patient monitoring scenario from the section “Four UC Settings”, the resurrecting duckling security policy works for secure transient pairing of a heartbeat monitor with a patient’s record. Here, the record takes the role of the master and the heartbeat monitor the role of the slave. Being at first in an **imprintable** state, as soon as the heartbeat monitor is attached to the patient, it is also imprinted by the patient’s record. Being imprinted, only the master is able to read the heartbeat and store it in the record. As soon as the patient is fully recovered and leaves the hospital, the patient’s record sends a kill command to the heartbeat monitor. The pairing is broken and the monitor is available for another patient.

### Out-of-Band Channels

Many communication technologies are found in the UC world that are suitable for cryptographic out-of-band transmission. We now give a non-exhaustive list of examples that have been already used in practice. Usually switching to a secondary channel requires some human intervention, but may also be triggered automatically by an appropriate protocol command on the primary channel. McCune, Perrig, and Reiter (2005) provide
a comparison of different out-of-band channels with respect to their security, convenience, and hardware requirements.

• **Infrared light** is used in IrDA interfaces for data transmission. It requires a direct line of sight and distances of 1 meter or less between the communicating parties. The salesman from the first scenario may use IrDA to establish a secure connection to a printer by exchanging a short-lived symmetric session key for the next print job or alternatively go through a pre-authentication protocol based on public key cryptography (like in Balfanz, Smetters, Stewart, & Wong (2002)). The idea of an IrDA-based “magic wand” that can be used to point to a device for establishing a secure connection is described by Spahic, Kreutzer, Kähmer, and Chandratilleke (2005).

• Dynamically generated **2D barcodes** can encode about 50-100 bits of information. They are suitable for modern mobile phones, which have high-resolution displays and built-in cameras. Claycomb and Shin (2006) have recently shown how to use colorized barcodes to exchange a public key fingerprint with an application to secure communications with ATMs. The authentication of an 802.11 access point using a sticker with a barcode on its casing is described in McCune, Perrig, and Reiter (2005).

• Devices that have a speaker and a microphone could use **audio** to transfer a fingerprint modulated on a piece of music. While the diffusion of the signal is larger than in the case of infrared, the presence of a second audio sender can be detected by a human. This technology is also suitable for **location-based authentication** of devices. An example is a café rewarding customers with media downloads by distributing access keys for the service via music played inside the building (Coulouris, Dollimore, & Kindberg, 2005).

• Buhan, Doumen, Hartel, and Veldhuis (2006) propose a user-friendly, two-party key agreement protocol that takes as an input **biometric data** of the device owners obtained via a grip pattern sensor. A downside of this approach, however, is the fact that biometric templates are static information and become useless once they fall into an opponent’s hands.

• Capkun and Cagalj (2006) apply **ultrasonic** to measure the distance and spatial direction of UC communication partners. Their concept of **integrity regions** forces the attacker to be physically close to the targets. When devices are operated by humans and the maximum distance is appropriately short, say 1 meter or less, the presence of a man-in-the-middle can be ruled out. Mayrhofer and Gellersen (2007) show how to transmit short messages over ultrasound in such a way that only a recipient at an expected distance from the sender can decode the signal.

• UC devices that do not support any of the above mechanisms, but have at least a **user interface**, can let the user enter a sequence of numbers by hand. Bluetooth applies an algorithm called “pairing” that allows two devices to agree on a common secret key derived from a random value (transmitted in the clear) and a static PIN (typically only a few characters long) that has to be known to each device.

• Even a simple user interface consisting, for instance, of an LED and a pushbutton is enough to transmit 1 bit of information telling the device that it is now safe to exchange public keys via the primary channel. This may happen when the devices are connected in an environment that is considered safe, for example at the user’s home.
The last example is an application of weak authentication as described by Arkko and Nikander (2002). According to the temporal separation principle, at a certain point in time the parties consider safe they exchange a key via the primary channel—without further protection, that is unauthenticated and in plaintext. Provided that no attacker is present in this phase, a long-term secure communication channel can be set up. The weak authentication mechanism completely defeats passive attacks (by exchanging public keys in an authentic fashion). It also raises the bar for active attacks, since a man in the middle who engaged during key agreement has to intercept and re-encrypt each subsequent message in order to remain undetected. A similar mechanism is used in the Secure Shell (SSH) protocol.

**RFID Security from EPC to E-Passport**

In this section the security of RFID systems is discussed, mainly by means of two examples: tags with a 96 bit Electronic Product Code (EPC) are used for the lifecycle management of commercial goods as described in Scenario 4. Such tags are at the lower end of the scale since they offer very limited functionality. At the other extreme, tags embedded in modern identity cards or e-passports (electronic passports) support state-of-the-art cryptographic algorithms.

The threat of abusive tracking of UC systems in general and RFID tags in particular has already been mentioned. Basically, tracking is what EPC tags are built for: Unlike barcodes, they do not simply indicate the manufacturer and product type, but a serial number that is distinct for each object. As access control and reader authentication is only an option in the EPC standards (see http://www.epcglobalinc.org), tags are inherently prone to skimming. That means that unauthorized parties reasonably close may read the chip without its owner noticing it. In the case of EPC tags that operate at a frequency of around 900 MHz, the specified reading range is 3 meters or more. E-passports use so-called proximity tags operating at a frequency of 13.56 MHz and a reading range of at most 15 cm. Note that fraudulent RFID readers do not necessarily adhere to standards and may therefore exceed reading ranges significantly (see e.g., Kirschenbaum & Wool, 2006).

**Countermeasures Against Skimming and Tracking**

EPC tags illustrate a common limitation of most RFID chips: Their identifier is defined during the production process and cannot be changed afterwards. This allows a tagged object to be misused for tracking a person it is linked to. In the discussion of anonymity in the section “Privacy-Enhancing Technologies”, we have already seen that identification is a problem affecting all network layers. RFID systems are no different in that respect as tracking may in theory happen on the application, the data link, or the physical layer. Consider, for instance, tags used in access control systems like the Hitag, Legic, or Mifare series. In order to prevent cloning and replay attacks, such tags may encrypt their communication on the application level. While this hides personal information of their owner, the tags themselves can be easily traced using their static data link layer ID.

In the EPC system, there is at least a basic, yet very effective privacy protection mechanism: A reader may send a tag the so-called “kill” command rendering it irreversibly inoperable. In order to prevent DoS attacks, the kill command requires authentication via a secret 32-bit number. The METRO group’s vision of an RFID-based store (see http://www.future-store.org) already encompasses a dedicated device behind the cash desk to kill tags. However, this comes at the price of excluding the tag ID from being used by the consumer at her home or for the purpose of recycling.
There are several non-destructive countermeasures against skimming and tracking, for a comparison see for example, Weis, Sarma, Rivest, and Engels (2003) or Juels, Rivest, and Szydlo (2003). For instance, RFID chips in identity cards can be shielded in a Faraday cage made of metal foil which is integrated in a wallet. “Clipped tags” (Karjoth & Moskowitz, 2005) provide a user-friendly way to deactivate a tag by manually separating chip and antenna. Such a method provides immediate visual feedback about the tag’s state. Instead of completely deactivating the tag, the antenna may be transformed in a way that reduces reading range considerably. Blocker tags (Juels et al., 2003) are devices that fool readers by simulating a pile of tags in order to prevent them to single out and communicate with a single one. As the EPC addressing scheme is hierarchical, a blocker tag may restrict the range of IDs selectively.

E-Passport Security Measures

Compared to EPC tags, the tags embedded in e-passports contain information that is much more sensitive. The specifications (ICAO, 2004a, 2004b) for “machine readable travel documents” provided by the International Civil Aviation Organization (ICAO), a United Nations body, govern the storage of personal and especially biometric data on the chip. Mandatory pieces of information include name, date of birth, nationality, and a digitized portrait photograph. While the standard defines iris scans and fingerprints as optional features, the member states of the European Union have agreed on including fingerprints in passports issued to their citizens (European Union, 2006).

As a baseline security mechanism, all e-passports have to implement “passive authentication”. This means that the data stored on the chip is digitally signed by the issuing nation. The signature algorithm may be one of RSA, DSA, or EC-DSA. While passive authentication prevents modifications of the chip, it does not help against chip cloning; only the optional feature “active authentication” does. Here, the chip is given a key pair the public key of which is included in the signed data set. During a challenge-response protocol, the chip proves possession of the private key and therefore its authenticity to the reader. While other chip data can be read and cloned, the private key is not disclosed similarly to a smartcard.

The ICAO specifications offer “Basic Access Control” (BAC) as an optional protection against eavesdropping, skimming, and tracking. When implemented on the chip, a reader is required to authenticate before being able to read non-sensitive data, that is, any information except fingerprints (and iris scans if present) which require additional authentication. EU e-passports implement BAC while US e-passports do not. This opens the door for “American-sniffing bombs” or even personalized RFID-enabled bombs (Juels, Molnar, & Wagner, 2005). The design of the BAC mechanism is guided by the idea that a passport has to be handed over to border control personnel, that is, physical access to the document is a prerequisite to reading the RFID chip. In order to start the communication, a reader must pass a challenge-response protocol with the chip. In this protocol, the reader has to implicitly prove that it knows two cryptographic keys derived from the passport number, the bearer’s date of birth, and the expiration date. These parameters are part of the so-called “Machine Readable Zone” (MRZ) which is why an e-passport has to be swiped through an optical reader first. The keys derived from the MRZ are used for 3DES encryption and integrity protection respectively. Note that, although the 3DES encryption uses a 112-bit key, the entropy obtained from the MRZ is much lower. The number of possible values can be at most \(2^{35}\) in theory, but practical figures may be significantly lower: The feasibility of a brute-force attack on Dutch e-passports, which provide encryption strength of around 35 bit due to their predictable serial numbers, is described
Security for Ubiquitous Computing

by Smith (2006). Note that even strong encryption does not prevent tracking by parties that have read the MRZ once, as the derived keys are static.

Another challenge-response scheme is used to protect the fingerprints on the chip. In this protocol, the chip authenticates the reader in order to decide whether to disclose the fingerprint or not. This mechanism involves national authorities that issue certificates to border control organization of foreign countries upon condition that fingerprints are treated in accordance with data security and protection legislation. These organizations in turn issue certificates to individual readers. A detailed description of the infrastructures used for issuing certificates to passport manufacturers and to readers respectively is given by Straub, Hartl, and Ruppert (2006).

RESEARCH OUTLOOK

This chapter gave an overview of the field of UC security taking into account UC characteristics and limitations. By means of representative settings, ranging from mobile work to RFID-based real-time enterprises, common security challenges were exhibited. At the heart of the chapter was the presentation of a bunch of concrete solutions to such challenges. This list is not claimed to be exhaustive, but it is an initial step toward a taxonomy. The solutions stand pars pro toto for UC security, as they are generic and scalable enough to be adaptable to a large number of scenarios.

Social acceptance of UC technology strongly depends on the security of UC systems. On the one hand, the growing use of UC in everyday life must not lead to security breaches as demonstrated, for instance, in a recent attack on RFID-enabled credit cards (Heydt-Benjamin et al., 2007). It is also a safe bet that conflicts of interest will again occur between privacy and law enforcement as we have already seen in the aftermath of 9/11, for example, concerning the traceability capabilities of UC. We hope that this chapter motivates the reader to delve deeper into the subject and refer to the following textbooks:

• IT security in general: Eckert (2006), Stallings (2006),
• Focus on cryptographic aspects: Buchmann (2004), Stallings (2005); Menezes, Oorschot, and Vanstone (1996),
• UC security in particular: Stajano (2002),
• RFID: Finkenzeller (2003).

Among the promising areas of UC security research we could not discuss in detail here are UC and its relation to Grid Computing (Storz, Friday, & Davies, 2003) and relay attacks on RFID systems (Kfir & Wool, 2005; Hancke 2006).

REFERENCES


A. Boukerche (Ed.), *Handbook of algorithms for wireless networking and mobile computing* (pp. 393–411). Chapman & Hall/CRC.


ternational Parallel and Distributed Processing Symposium. IEEE Computer Society Press.


ADDITIONAL READING


*This work was previously published in the Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises, edited by M. Mühlhäuser and I. Gurevych, pp. 337-362, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Section II
Development and Design Methodologies

This section provides in-depth coverage of conceptual architectures, frameworks and methodologies related to the design and implementation of interfaces and technologies that promote successful human-computer interaction. Throughout these contributions, research fundamentals in the discipline are presented and discussed. From broad examinations to specific discussions on particular frameworks and infrastructures, the research found within this section spans the discipline while also offering detailed, specific discussions. Basic designs, as well as abstract developments, are explained within these chapters, and frameworks for designing and personalizing interfaces for medical equipment, televisions, and automobiles are discussed.
Chapter 2.1

Measuring the Human Element in Complex Technologies

Niamh McNamara
University College Cork, Ireland

Jurek Kirakowski
University College Cork, Ireland

ABSTRACT

Measuring satisfaction can provide developers with valuable insight into the usability of a product as perceived by the user. Although such measures are typically included in usability evaluations, it is clear that the concept itself is under-developed. The literature reveals a lack of cumulative, systematic research and consequently the field is in “disarray” (Hornbæk, 2006). Clearly, the area needs a strong theoretical foundation on which to base research. This paper reviews the literature on user satisfaction and proposes a conceptualisation and definition of the concept that will aid researchers in the development of valid measures.

USER SATISFACTION AS A USABILITY PARAMETER

The ISO 9241-11 standard suggests assessing usability in terms of performance by measuring effectiveness and efficiency, and in terms of internal state by measuring satisfaction. It recommends that at least one measure of each parameter should be included in usability evaluations. Despite this, Lindgaard and Dudek (2003, p. 430) have observed that the:

Human-computer interaction literature has traditionally concerned itself almost exclusively with effectiveness and efficiency with satisfaction being regarded (as) mainly a by-product of great usability.

Although satisfaction is a central concept in HCI it has been described as “underdeveloped” (Tractinsky, 2004, p. 352). It is the criterion most often included in usability evaluations (Kirakowski, 2002) but there is little understanding on the part of evaluators of what exactly they are measuring, and often the tools they employ are poorly developed. This article will illustrate the consequences of the lack of a cumulative, systematic approach to research that is evident in the
literature on user satisfaction (Hornbæk, 2006). It will be shown that much of what is known about satisfaction is based on relatively little empirical work and represents opinion rather than fact. The diversity of adjectives used in satisfaction measures is symptomatic of the lack of a cumulative, systematic approach to research in the field. Hornbæk (2006, pp. 90, 91) observes that “comparisons across studies become difficult because of this diversity, as does the understanding of distinct dimensions of subjective satisfaction.”

Clearly, the field needs a theoretical foundation on which to base research beginning with a sound conceptualisation of the construct and a suitable definition.

There are very few definitions of satisfaction to be found in the HCI literature. Possibly the closest the field has to a standard definition is that provided by ISO 9241-11, where satisfaction is defined as “freedom from discomfort and positive attitudes towards the product.” However, this is quite a broad definition that gives little insight into the nature of the satisfaction construct. Although there is no widely-accepted definition, there seems to be a number of common elements that most researchers agree on. Satisfaction is:

1. A response (either cognitive or emotional) on the part of the individual (Chin, Diehl, & Norman, 1988; LaLomia & Sidowski, 1990; Lindgaard & Dudek, 2003)
2. To an interactive system
3. Based on the degree to which the user feels the device facilitates the completion of a particular goal (LaLomia & Sidowski, 1990; Bevan & MacLeod, 1994; Kirakowski, 1999)
4. That is linked to usage (Lund, 2001; Kirakowski, 2002)

These characteristics are similar to those of the attitude construct in social psychology. Indeed, satisfaction is often described in the literature as an attitude (Melone, 1990; Nielsen, 1993; Keinonen, 1998; Shneiderman, 1998). Melone (1990) suggests that conceptualising satisfaction in this way provides the necessary theoretical foundation for research and “retains the essential elements embodied in the user satisfaction construct.”

This article starts by reviewing attempts to measure and uncover the structure of user satisfaction from early HCI, research on Information Systems, and from recent HCI. Checklist approaches to usability in general (e.g., Reiterer & Oppermann, 1993; Ravden & Johnston, 1989) will not be reviewed here as they are too broad in scope; neither will the influential research of Dzida, Herda, and Itzelfeldt (1978) for the same reason. Although each of these approaches creates methods of assessment which involve questionnaires, they only look superficially like satisfaction questionnaires; they are closer to checklists.

The measurement of satisfaction would greatly benefit from the provision of a sound theoretical base for research (Melone, 1990; Hornbæk, 2006). Melone (1990) suggests that this would be best achieved by conceptualising satisfaction as an attitude. Although HCI uses the term attitude regularly in discussions on satisfaction, rarely are the implications of this way of conceptualising satisfaction addressed. This article will go on to illustrate how this conceptualisation does indeed retain satisfaction’s essential elements while also clarifying some theoretical issues facing researchers, specifically (1) determining the nature of the user’s response (i.e., whether it is cognitive or emotional), (2) conceptualising satisfaction as a characteristic of the interaction (quality in use) rather than as a product attribute (quality of features) and (3) understanding the relationship between satisfaction and usage behaviour. Following conceptualisation of satisfaction as an attitude, a definition of the construct in keeping with its character as a usability parameter is offered.

Finally, four proposals are made for the future development of measuring instruments for satisfaction.
UNCOVERING THE STRUCTURE OF USER SATISFACTION

Early Satisfaction Research in HCI

Most of the research on satisfaction was carried out in the context of developing questionnaires to measure user satisfaction during the late 1970s and 1980s. At this time, developers were becoming more aware of the need to evaluate the usability of their products. Various types of questionnaires were developed. Some were single-use instruments constructed for specific studies that often only provided a global measure of satisfaction, while others were developed in industry, such as Brooke’s (1996) system usability scale (which was developed in the 1980s while the author was employed by the Digital Equipment Corporation, DEC) and Lewis’ (1995) IBM usability questionnaires. While instruments like SUS and the IBM questionnaires have proven to be quite useful when evaluating software, they have offered little in the way of progress in understanding the nature of user satisfaction. For this, one has to turn to instruments developed in an academic context. While few such instruments exist, they are worth examining primarily for their efforts to uncover the structure of user satisfaction, and also as a potential starting point for research on user satisfaction with new technology.

The questionnaire for user interface satisfaction (QUIS) (Chin, Diehl, & Norman, 1988) and the computer user satisfaction inventory (CUSI) (Kirakowski & Dillon, 1988) are two early satisfaction questionnaires developed at roughly the same time, but that conceptualised the construct in two different ways.

QUIS, which is still widely used today, was originally constructed to measure “the user’s subjective rating of the human-computer interface” (Chin et al., 1988, p. 213). The authors viewed satisfaction as dependent on system features, and respondents were asked to report their satisfaction with the use of these features on a 9-point scale. The questionnaire had five sections referring to different aspects of the system, namely, Overall Reactions, Screen, Terminology and System Information, Learning, and System Capabilities. Chin et al (1988) carried out a factor analysis on the responses to version 5 of the questionnaire from which they extracted the following four factors: (1) Learning, (2) Terminology and Information Flow, (3) System Output and (4) System Characteristics. Carrying out factor analyses provides an important insight into the nature of a construct, however Chin et al. (1988) did not provide a comprehensive explanation of what the four factors measured. Although they attempted to uncover the underlying structure of satisfaction, the authors failed to provide a meaningful conceptualisation of the construct.

This is in contrast to the development of the Computer User Satisfaction Inventory (CUSI, Kirakowski & Dillon, 1988). This research was informed by Bandura’s (1977) Self-efficacy Theory. Self-efficacy refers to the individual’s belief that he or she can accomplish a particular task. The authors hypothesised that satisfaction could be measured on two dimensions: Affect, which assesses the “(individual’s) feelings of fear or pleasure” and Competence, which assesses the “individual’s feelings of mastery over the computer system” (Kirakowski & Corbett, 1988, p.330). Kirakowski and Dillon (1988) not only created a questionnaire that could be used as an alternative to other more resource- and labour-intensive methods, but also attempted to provide a conceptual understanding of the satisfaction construct.

CUSI includes an affective dimension that is not present in QUIS. While CUSI’s developers did not provide a definition of satisfaction per se, it is clear that they did believe it to contain elements relating to user emotions (Affect) and performance (Competence). This would appear consistent with later definitions of satisfaction such as ISO 9241-11, Bevan and MacLeod (1994) and Bevan and Azuma (1997).
The conceptualisation of usability as *quality in use* in the 1990s greatly influenced research on user satisfaction. Satisfaction was no longer widely regarded as a product property, but as a property of the interaction between the user and the product (Bevan, Kirakowski, & Maissel, 1991). This proved to be a much more successful approach to measuring satisfaction than the previous quality of features approach espoused by Chin et al (1988). Their focus on specific attributes of the product resulted in model instability and limited applicability across applications for questionnaires that followed this route. Thus, attention focused on constructing models of satisfaction that “(attempted) to measure the way the software being rated fulfils some latent need within the user community” (Kirakowski, 1999, p.129).

Examining user reactions to an interactive system applies a psychological approach to the investigation of user satisfaction. The *software usability measurement inventory* (SUMI) (Kirakowski & Corbett, 1993; Porteous, Kirakowski, & Corbett, 1995; Kirakowski, 1996), based on this approach to user satisfaction measurement, has become the *de facto* instrument for evaluating software applications (Kirakowski, 2002). It measures “user-perceived software quality,” which it is hypothesised, consists of five dimensions: Affect, Controllability, Efficiency, Helpfulness and Learnability.

This model has also been successfully applied to the evaluation of user satisfaction with Web sites in the *Website Analysis and MeasureMent Inventory* (WAMMI) (Kirakowski, Claridge, & Whitehand, 1998), illustrating its applicability across technological domains. The five factors of the model were modified slightly for Web sites—the main difference being that the Affect dimension is replaced by an Attractiveness dimension.

We now turn to research on satisfaction from the discipline of management information systems (MIS). While this literature is often included in reviews of user satisfaction research, rarely is it noted that the constructs of *user satisfaction* and *user information satisfaction* are entirely different. The following section will explain these differences and advise caution in borrowing concepts across disciplines.

### Information Systems Research on Satisfaction

HCI and management information systems (MIS) are two disciplines that have a significant interest in assessing satisfaction with technology, and consequently, often cite each other’s research. However, researchers from both disciplines do not seem to recognise the vastly different perspectives each takes on what they perceive to be the same construct, particularly concerning the role of usability in operationalisations and definitions of satisfaction.

The term *user information satisfaction* is itself suggestive of the differences between the ways satisfaction is conceptualised in the HCI and MIS disciplines. MIS is primarily interested in the benefits of technology for decision-making in an organisation and its impact on productivity. The goal of HCI, on the other hand, is to create technology that supports users (Preece, Rogers, Sharp, & Benyon, 1994). The definitions of satisfaction from MIS are broadly similar to ISO 9241-11’s definition of satisfaction. For example, Bailey & Pearson (1983) define it as “the sum of feelings or attitudes, both positive and negative, toward the spectrum of factors affecting a specific situation” (p.531). Similarly, Doll & Torkzadeh (1988) define satisfaction as “the affective attitude towards a specific computer application by someone who interacts with the application directly” (p.261). However, the operationalisations of the construct between MIS and HCI are quite different. Typically, the MIS literature treats usability as one dimension of user satisfaction, while HCI treats satisfaction as a usability parameter. The user information satisfaction construct is influenced by user role rather than conceptualisations of...
usability, while the user satisfaction construct is typically influenced by both. This may be the reason why the usability or ease of use dimension was only included in (MIS) user information satisfaction questionnaires with the advent of end-user computing.

The computing environment and the role of the end-user were undergoing rapid changes during the late 1980s, and with this came a change in the way satisfaction was conceptualised in the MIS literature. Early studies such as those conducted by Bailey & Pearson (1983) and Ives, Olson, and Baroudi (1983) dealt with a system that included the computer, the end-user, and the programmer/analyst. As the end-user became more autonomous, attention was increasingly focused solely on the application, and the output and usability began to be considered. Organisational structures such as EDP personnel were not included in evaluations as prominently as had been the case previously, although some evaluations still included the adequacy of the technical support provided.

A new definition of satisfaction in accordance with these changes was proposed by Doll and Torkzadeh (1988). They defined end-user computing satisfaction as “the affective attitude towards a specific computer application by someone who interacts with the application directly” (p. 261). Doll and Torkzadeh (1988) argued that because the end-user was now interacting with the system directly instead of through programmers or analysts, there was a need to consider the usability (ease of use) of the application. The proposed model, which the authors claimed was more suitable for an end-user computing environment than previous efforts, consisted of five factors, namely Content, Accuracy, Format, Ease of Use, and Timeliness.

Doll and Torkzadeh (1988) did include items referring to ease of use that might lead one to infer that this conceptualisation of end-user computing satisfaction is closer to HCI’s user satisfaction than user information satisfaction. However, this is not the case. Doll & Torkzadeh (1988) still focus on the information product, that is, satisfaction with the information the system provides to the user to assist in decision-making. Furthermore, it seems to oversimplify the concept of usability by simply including the following two items to measure it: “is the system user friendly?” and “is the system easy to use?” (Doll & Torkzadeh, 1988, p. 268). Clearly, these items do not reflect the complexity of the usability construct.

Doll and Torkzadeh’s (1988) end-user computing satisfaction measure became the standard instrument within the MIS community. Much of the research that was conducted after the publication of this questionnaire either involved further validation studies of the instrument or attempts to investigate the relationship between satisfaction and other variables such as usage, user involvement in design, and so on. The research focus shifted from purely measurement issues to theoretical issues, particularly the need to identify the nature of the satisfaction construct and the importance of establishing a sound theoretical base for research. Thus, while the satisfaction construct differs between the two disciplines of MIS and HCI, one thing HCI can and should learn from MIS is the importance of providing a theoretical foundation for research (Melone, 1990).

Recent Approaches to Satisfaction in HCI

Towards the late 1990’s, little further progress had been made in HCI in revealing the true nature of the satisfaction construct. At this time, researchers were becomingly increasingly interested in investigating the wider user experience. There are two emerging approaches to satisfaction research in light of this development.

In the first approach, researchers believe that the satisfaction construct can encompass the wider elements of the user experience (e.g., Han & Hong, 2003; Lindgaard & Dudek, 2003; Yun, Han, Hong, & Kim, 2003), particularly with the
inclusion of new dimensions. The operationalisations proposed by these researchers are very different to those already discussed. For example, Lindgaard and Dudek (2003, p. 430) define satisfaction as an “expression of affect” and suggest it consists of four dimensions: Aesthetics, Likeability, Expectation, & Emotion. Similarly, Yun et al. (2003) define satisfaction as the user’s subjective impression of the device which is informed by performance and emotional aspects. The dimensions of satisfaction they propose include Luxuriousness, Simplicity, Attractiveness, Colourfulness, Texture, Delicacy, Harmoniousness, Salience, Ruggedness, and Overall Satisfaction. Finally, Han and Hong (2003) define satisfaction as meaning that “products should provide pleasure and/or attractiveness” (p.1441). The corresponding dimensions they propose include Volume, Shape, Elegance, Harmoniousness, Simplicity, Comfort, Attractiveness, and Overall Satisfaction. Clearly, these conceptualisations of satisfaction have little to do with perceived usability; rather they refer to aesthetic concerns.

The second approach taken by researchers involves a critique of the relevance of satisfaction. It is argued that the construct should be replaced (Jordan, 1998, 2000; Dillon, 2001). Just like the call to go “beyond usability” (Logan, 1994), Dillon (2001) called for the creation of a construct that goes “beyond satisfaction.” This is because he believed that the concept of satisfaction as it had been understood up to then was inadequate for application to new technology. He proposed the Affect construct as a replacement which is predominantly concerned with user emotions.

Finally, recent approaches to satisfaction are moving from a view of the construct as a characteristic of the interaction to one that conceptualises satisfaction as an intrinsic product attribute. Even though satisfaction has been shown not to be a product characteristic (Kirakowski, 1999), new conceptualisations of satisfaction proposed by Yun et al. (2003) and Han and Hong (2003) take an ergonomic approach and conceptualise satisfaction in terms of product properties. Yun et al. (2003) and Han & Hong (2003) argue that satisfaction is dependent on design. Thus, each of their dimensions consists of particular product properties known as human interface elements (HIEs) that they believe contribute to satisfaction. Using the method of Kansei Engineering, they maintain that it is possible to predict a user’s level of satisfaction with a product before it is released. In doing this, Yun et al. (2003) and Han and Hong (2003) ignore the impact of the context of use on the interaction between the user and the device. Consequently, we see the same difficulties inherent in their approach that were encountered by others who took an ergonomic approach to defining and measuring usability. Before any evaluation could take place using this approach, evaluators would have to ascertain which particular dimensions of satisfaction and which HIEs were relevant to the device under evaluation, making it a time-consuming method, and one that is virtually impossible to apply across devices.

It seems that some researchers are keen to align satisfaction with the concept of experience for no other reason except that it is the only “subjective” construct consistently referred to in the usability literature (Dillon, 2001). However, there is a clear distinction, which is often not made, between the concepts of usability and experience. (This issue is outside the scope of this article; see McNamara and Kirakowski (2005, 2006) for further discussion.)

The satisfaction construct is becoming increasingly broad as is evident in the new definitions and operationalisations of satisfaction that have been proposed in the literature. Indeed, the only common link remaining between recent conceptualisations of satisfaction and earlier ones seems to be that it is still viewed as a multidimensional construct. Unfortunately, there is little theoretical or empirical justification offered as the basis for these new models of satisfaction. Indeed many “new” dimensions appear to be borrowed from other disciplines such as Consumer Psychology.
The absence of clear content boundaries has led to this indiscriminate borrowing of concepts and theories across disciplines.

**What is Satisfaction?**

It is clear from this review of the user satisfaction literature that there are three theoretical challenges facing researchers.

First, little insight is given into the nature of the construct. Some only offer broad definitions of satisfaction and little else (e.g., ISO 9241-11). Others offer both a definition and model of the construct. But in many cases the model does not seem to correspond to the definition. For example, Yun et al. (2003) define satisfaction in terms of both performance and affective elements, yet their operationalisation of the construct does not include any dimensions relating to user performance. Some view the user response as exclusively cognitive (e.g., Chin et al., 1988), others as exclusively emotional (e.g., Han & Hong, 2003; Lindgaard & Dudek, 2003), and still others who take a combined approach including both cognitive and affective elements in their models of satisfaction (e.g. Porteous et al., 1995; Kirakowski et al., 1998).

Second, some researchers view satisfaction as a characteristic of the interaction between the user and the product (e.g. Porteous et al., 1995; Kirakowski et al., 1998) which is in keeping with the quality in use approach to usability proposed in ISO 9241-11 and widely accepted by the majority of usability professionals. However, others view satisfaction as a product attribute (e.g., Chin et al., 1988; Han & Hong, 2003; Yun et al., 2003) indicating a return to the generally discredited quality of features approach to measurement.

Finally, it appears to be taken for granted that satisfaction is linked to usage. Many mention this link (e.g., Lund, 2001; Kirakowski, 2002), but there is little evidence of any investigations (in the HCI literature at least) into the nature of this relationship.

It is unlikely that these issues can be resolved without better theoretical foundations for research. It is argued that conceptualising satisfaction as an attitude can provide such a base (Melone, 1990).

**USER SATISFACTION AS AN ATTITUDE**

The study of attitudes, their structure, formation, and their relationships with other constructs (most notably behaviour) has occupied a central position within Social Psychology almost since its inception (Palmerino, Langer, & McGillis, 1984; Ajzen, 2001).

Although there does not exist a standard definition of attitude (Fishbein & Ajzen, 1975; Dawes & Smith, 1985; Olson & Zanna, 1992), it is clear that evaluation is an important component. This is illustrated in the following four definitions of attitude taken from classics in the attitude literature: (1) “(A term) used to refer to a relatively enduring tendency to respond to someone or something in a way that reflects a positive or negative evaluation of that person or thing” (Manstead, 1996, p. 3); (2) “(a tendency) to evaluate an entity with some degree of favour or disfavour, ordinarily expressed in cognitive, affective, and behavioural responses” (Eagly & Chaiken, 1993, p.155); (3) “a person’s feelings toward and evaluation of some object, person, issue, or event” (Fishbein & Ajzen, 1975, p.12) and finally, (4) “a summary evaluation of a psychological object captured in such attribute dimensions as good-bad, harmful-beneficial, pleasant-unpleasant, and likable-unlikable” (Ajzen, 2001, p. 28).

These definitions demonstrate that theorists define attitudes in affective (e.g., Fishbein & Ajzen, 1975), behavioural (e.g., Manstead, 1996), and cognitive (e.g., Ajzen, 2001) terms. The definition offered by Eagly and Chaiken (1993) where attitudes are defined in terms of all three components (affect, behaviour, and cognition)
represents the most common conceptualisation of attitude structure. Affect refers to the individual’s emotional response to the target of the attitude. Behaviour is the action engaged in by the individual towards the target, which is assumed to reflect the attitude held by the individual. Cognition refers to the knowledge and beliefs held by the individual that link the target to particular attributes.

Definitions of attitude vary because theorists disagree regarding the relative importance of the various processes of affect, behaviour, and cognition involved in attitude formation.

Explaining attitude formation on the basis of affective processes is based primarily on two areas of research: the principles of conditioning, both operant and classical; and, on Zajonc’s (1980) research on the Mere exposure effect. Classical conditioning involves pairing an object with a stimulus that elicits an affective response. Through repeated association, the attitude object comes to elicit the affective response and an attitude is formed. Operant conditioning typically involves reinforcement of positive or negative statements by an individual. The attitude that subsequently forms depends on whether positive or negative statements were reinforced. The mere exposure theory states that familiarity with objects, based on the frequency with which they are encountered, increases liking for those objects.

Bem’s (1972) Self-perception theory is often used to describe attitude formation in terms of behavioural processes. According to this theory, individuals infer their attitudes from their behaviour, that is, engaging in a particular behaviour towards an object (reflecting favour or disfavour) is indicative of their underlying attitude. Bem (1972) states that individuals take context into account when interpreting their own behaviour. Behaviour is seen to be reflective of an underlying attitude only if there are no forces external to the individual that might be inferred to be causing his or her behaviour.

Finally, according to the cognitive perspective, individuals gain information about an object through direct or indirect experience with it. This information results in the formation of beliefs about the object which links it to particular attributes. If the beliefs are mainly favourable, the resulting attitude will be positive, and if the beliefs are mostly unfavourable, a negative attitude towards the object will result.

It seems to make sense to consider satisfaction as an evaluative state that can be formed through any of the affective, behavioural, or cognitive processes. Rather than broadening the construct beyond recognition, it is proposed here that it is best to define it (based on the definition of attitude proposed by Ajzen, 2001) as a summary evaluation of an interactive system captured in terms of how easy the user thinks it is to use. This definition is in keeping with the conceptualisation of satisfaction both as an attitude and as a usability parameter.

Now that a definition of satisfaction has been offered, the ways in which the attitude literature can inform three important theoretical issues facing satisfaction researchers will be discussed: (1) the nature of the user response (i.e. whether it is cognitive or affective-emotional); (2) whether this response is best described as a characteristic of the interaction (quality in use) or as an intrinsic product attribute (quality of features); and (3) understanding the relationship between satisfaction and usage.

The Nature of the User Response

As noted in the review of the user satisfaction literature, some view satisfaction as a cognitive response and others as an emotional response. A minority take a combined view. Consequently, there is a debate as to which view is the correct one. Consumer Psychology typically views satisfaction as an emotional response (Giese & Cote, 2000), and given that this literature has influenced recent operationalisations of user satisfaction, it is no surprise that definitions proposed by Lindgaard and Dudek (2003) and Han and Hong
Measuring the Human Element in Complex Technologies

(2003) have promoted the view of satisfaction as an exclusively emotional response to a product. Some attitude literature in psychology suggests that attitudes can be formed on the basis of emotional processes. Thus, these researchers could point to this literature as providing support for their conceptualisation of satisfaction. However, attitude theorists also do make a clear distinction between affect and emotion, which suggests that satisfaction (or any attitude), cannot be described as simply an emotional response tout court.

There are two ways in which affect, as it is conceptualised by attitude researchers, can be distinguished from general emotions and moods (Judd & Johnson, 1984). First, affect relates to a specific object (Judd & Johnson, 1984; Eiser, 1994). Second, the valence of affect is generally conceptualised as being unidimensional, that is, either positive or negative, while “the more general study of emotions…is full of attempts to invoke more richly defined distinctions” (Judd & Johnson, 1984, p. 66).

This seems to clarify the distinction between viewing satisfaction as an emotional response and viewing it as an attitude that can be formed on the basis of affective processes. However, it does not resolve the issue of which set of processes are primarily responsible for the attitude held by an individual. Is it the affective processes (i.e., the feelings the object elicits) or the cognitive processes (i.e., beliefs about that object)?

Evidence supporting both sides of the debate can be found in the attitude literature. Proponents of the affective perspective argue (in response to the dominant cognitive perspective on attitudes) that affect is more important in attitude formation than cognition because affective aspects are more accessible in memory than cognitive aspects (Ajzen, 2001). Their position is said to be further strengthened by research which states that if there is disagreement between beliefs and feelings, (i.e., if they are of opposite valence), feelings take precedence (Ajzen, 2001).

Proponents of the cognitive perspective, in turn, argue that cognition (specifically awareness) mediates affective processes. Thus, cognitive processes are more influential in forming attitudes than affective processes. Researchers argue that operant conditioning effects are greater when experimental participants are aware of a contingency between responses and reinforcement (Eagly & Chaiken, 1993). Effects of this type are also seen in studies on classical conditioning. Such effects are also seen in research on the mere exposure effect in that participants may be aware that they should respond in a particular way to stimuli that occur more frequently than others. However, there is research involving subliminal presentation of stimuli that produced both conditioning and mere exposure effects. This suggests that conscious stimulus recognition is not necessary for positive attitudes to form toward objects (Eagly & Chaiken, 1993).

Eagly and Chaiken (1993, p. 421) note that there is “a lack of clear criteria to distinguish between affective and cognitive explanations” of attitude formation. For example, some learning theorists believe that it is not responses that are learned in instances of classical conditioning, but mental representations of relationships between stimuli and responses. The authors maintain it is somewhat misleading to refer to conditioning as involving purely affective mechanisms; cognitive processes, they say, can also be used to explain the effects of classical conditioning. According to the principles of classical conditioning, conditioned responses occur automatically in the presence of a conditioned stimulus, that is, they are not mediated by cognitive processes. However, Eagly and Chaiken (1993) argue that associative network models of memory can provide an alternative explanation for why these responses occur. This theory states that each object and each response is represented by a concept node and an evaluative node, respectively, in memory, and that these are linked by an associative bond. A conditioned stimulus (also represented by a concept node)
that is not originally linked with the response becomes linked with it through repeated association. A network is formed between all the concept and evaluative nodes. Activating any one node activates the entire network. It is in this manner that attitudes are formed. Clearly, when the same process can be explained using appeals to both affective and cognitive mechanisms, it is difficult to determine which explanation is correct.

Ajzen (2001, p. 29) notes, “The debate over the cognitive versus affective basis of attitudes has yet to be resolved.” So it seems most prudent to follow a research strategy in which attitudes are hypothesised to form as a result of both affective and cognitive processes (Judd & Johnson, 1984; Ajzen, 2001) rather than either one or the other. The relationship between these two components should be characterised as one of “reciprocal causality” (Judd & Johnson, 1984, p. 79). For example, objects that elicit more intense affect are more easily remembered and judged more quickly than those that elicit less intense affect. In addition, affective responses influence how the individual organises and stores information (Judd & Johnson, 1984).

The implication of this discussion for research on user satisfaction is that it is inappropriate to conceptualise the satisfaction construct as an “expression of affect” where affect denotes emotion only. However, conceptualising satisfaction as an attitude does not preclude the possibility of an affective component considering the reciprocal relationship between affect and cognition.

**Satisfaction: Quality in Use or Quality of Features?**

The second issue facing satisfaction researchers that could benefit from application of theories from the attitude literature is whether it is best to view the construct as a characteristic of the interaction between the user and the product or as a product attribute. As mentioned earlier, it is widely accepted that usability is best conceptualised as a characteristic of the interaction. Given that satisfaction is a usability parameter, one could assume that it also would be conceptualised in a similar manner. However, recent approaches to investigating the construct have proposed a return to the widely discredited quality of features approach. “Quality of features” implies that the level of user satisfaction is dependent on product features (e.g., Han & Hong, 2003; Yun et al., 2003). This approach, referred to very insightfully as “design reductionism” by Hassenzahl, Beu and Burmester (2001), is unlikely to prove successful. Desmet and Hekkert (2002, p. 67) assert that there is not a direct relationship between emotion and product design, and that “emotions are not elicited by product characteristics as such, but by construals based on these characteristics.”

A similar issue arose in the attitude literature. Research on attitudes often neglected the fact that attitudes are social constructs, and instead, they were generally conceptualised as individual attributes (Manstead, 1996). However, newer conceptualisations of attitude argue for the inclusion of context, a variable not often considered in the literature, but which has a significant bearing on attitudes (Palmerino, Langer, & McGillis, 1984; Ajzen, 2001). It is suggested that context may condition attitude (Ajzen, 2001.)

Palmerino et al. (1984, p. 181) define an attitude as “a relationship between two entities where one entity is a person and the other is a person or object” and point out that “the form of this relationship (i.e., the attitude) can be radically different given different contexts” (p.179).

The conceptualisation of attitude proposed by Palmerino et al. (1984) seems well suited to user satisfaction research for two reasons. First, because of the inclusion of context and second, because it views attitude as the relationship between an individual and a psychological object, rather than hypothesising that the attitude resides totally in the person or (absurdly) totally in the object. This bears a striking similarity to the notion of _quality in use_ where usability is conceptualised...
as a property of the interaction between the individual and the device under evaluation.

This discussion suggests that viewing satisfaction as a product characteristic is inadequate (which is what is implied by Yun et al., 2003, and Han & Hong, 2003). Attitude research studies provide the necessary theoretical basis for proposing that satisfaction is a characteristic of the interaction, and support an approach such as that followed by Porteous et al. (1995) which focuses on user reactions to the product rather than referring to specific product features.

**Investigating the Relationship between Satisfaction and Behaviour**

Satisfaction is considered to be an important indicator of future user behaviour (Kirakowski, 2002). However, there is little empirical evidence available to support this claim. Although Melone (1990) cautioned that attitudes only may play a moderating role in determining behaviour, researchers tend to oversimplify the relationship between satisfaction and usage.

It was thought that knowledge of an individual’s attitude could be used to predict his or her behaviour (see the discussion in Eiser, 1994). However, when studies by LaPiere (1934), DeFleur and Westie (1958), and Wicker (1969) found there to be little or no correlation between verbal expressions of attitude and overt behaviour, many called for the abandonment of research into the attitude construct altogether. Fishbein and Ajzen (1975) maintain that poor correlations between measures of attitude and overt behaviours were a consequence of the flawed attempt to predict specific behaviours using general measures of attitude. They developed their own attitude framework. This was an attempt to standardise the terminology used, and to clarify the relationships between the various hypothesised components of the attitude construct and the relationships of the attitude construct itself with other constructs. Fishbein and Ajzen (1975) maintain that the concepts of beliefs, attitude, intention, and behaviour have been used interchangeably in the literature, but by making distinctions between them, many of the inconsistencies found in research on attitudes can be reduced.

In their Theory of Reasoned Action, Fishbein and Ajzen (1975) contend that beliefs are the “fundamental building blocks” that determine attitudes. Intentions are a function of the beliefs the individual has about the behaviour in question, that is, the consequences of performing a specific behaviour and the individual’s evaluation of those consequences. Whether the individual performs the behaviour in question depends on his or her behavioural intentions, which in turn are determined by both the individual’s belief structure and what Fishbein and Ajzen (1975) term the subjective norm: “this refers to the extent to which you believe that your acting in a given way will earn the approval or disapproval of other people whose opinions you value” (Eiser, 1994, p. 21).

In the end, the Theory of Reasoned Action provides a good justification for proposing that satisfaction by itself may never be sufficient to predict behaviour. However, through the Affective route, attitudes must surely influence behaviour and, therefore, we will expect to see a broad degree of relationship between the two. The hypothesis is, therefore, that discrepancies between satisfaction and usage must be attributable to biasing contextual factors. It is not enough in any one case to show that there is a poor relationship between satisfaction and behaviour; all one needs for that are measuring instruments of poor discriminability. If a poor relationship is found, then one must hunt to find the biasing contextual factors.

**CONCLUSION**

The preceding discussion illustrates that Melone (1990) was correct in asserting that a broad conceptualisation of satisfaction as an attitude retains the construct’s “essential elements” while enabling
access to the necessary theoretical foundation for research in this area. The discussion has gone on to attempt to uncover what these essential elements are. It was argued that using theories and models from the attitude literature can help with some of the challenges faced by researchers, particularly in constructing definitions and operationalisations of satisfaction which may lead to scales of measurement.

The gist of this paper can be summarised in four statements.

First, it is proposed here that the construct should not be broadened as this changes the construct into unrecognisable forms. It will lead to greatest consensus to define satisfaction (based on Ajzen’s 2001 definition of attitude) as a summary evaluation of an interactive system captured in terms of how easy a user thinks a system is to use. For HCI, satisfaction as so defined is an element of the construct of Usability. Defining it in this manner ensures that satisfaction remains a measure of usability as *quality in use*, rather than a broader measure of the entire user experience.

Second, it is suggested that research into the structure of the satisfaction concept is important. This kind of research seeks to discover the main directions along which users are thinking when they interact with a particular class of products. It is likely that such thinking will change over time as technology is assimilated into the way we live. Structural information can be used formatively (in the development of technology), as well as summatively (in the assessment of technology.)

Third, it is proposed that satisfaction is the result of an evaluative state in the user that can be formed through affective, cognitive, or behavioural processes. These processes may be related through emotion, but satisfaction is not an exclusively emotional response on the part of the user toward a product.

Finally, it is suggested that in accordance with the conceptualisation of usability as *quality of use*, the level of user satisfaction with a product is dependent on the interaction between a user and a product, and the context in which that interaction takes place. Satisfaction should not be regarded as being solely a product attribute.

We propose that these are four fundamental principles in the development of scales of measurement of the concept of user satisfaction in HCI.

**REFERENCES**


Chapter 2.2
Human–Centric Evolutionary Systems in Design and Decision–Making

I. C. Parmee
University of the West of England, UK

J. R. Abraham
University of the West of England, UK

A. Machwe
University of the West of England, UK

ABSTRACT
The chapter introduces the concept of user-centric evolutionary design and decision-support systems, and positions them in terms of interactive evolutionary computing. Current research results provide two examples that illustrate differing degrees of user interaction in terms of subjective criteria evaluation; the extraction, processing, and presentation of high-quality information; and the associated improvement of machine-based problem representation. The first example relates to the inclusion of subjective aesthetic criteria to complement quantitative evaluation in the conceptual design of bridge structures. The second relates to the succinct graphical presentation of complex relationships between variable and objective space, and the manner in which this can support a better understanding of a problem domain. This improved understanding can contribute to the iterative improvement of initial machine-based representations. Both examples complement and add to earlier research relating to interactive evolutionary design systems.

INTRODUCTION
Uncertainty and poor problem definition are inherent features during the early stages of design and decision-making processes. Immediate requirements for relevant information to improve
understanding can be confounded by complex design representations comprising many interacting variable parameters. Design constraints and multiple objectives that defy complete quantitative representation and therefore require a degree of subjective user evaluation further inhibit meaningful progression. Machine-based problem representation may, initially, be based upon qualitative mental models arising from experiential knowledge, group discussion, and sparse available data. However, such representations, coupled with user intuition, play a significant role in defining initial direction for further investigation. Concepts based upon current understanding require both quantitative and qualitative exploration to generate relevant information that supports and enables meaningful progress.

The chapter presents research and development relating to powerful machine-based search and exploration systems that, through appropriate user interaction, allow both quantitative and qualitative evaluation of solutions and the extraction of information from complex, poorly understood design and decision-making domains. The integration and capture of user experiential knowledge within such systems in order to stimulate, support, and increase understanding is of particular interest. The objective is the realisation of user-centric intelligent systems that overcome initial lack of understanding and associated uncertainty, support an improving knowledge-base, allow the integration of subjective judgement, and stimulate innovation and creativity.

**INTERACTIVE EVOLUTIONARY COMPUTATION (IEC)**

Interactive evolutionary computing (Takagi, 1996) mainly relates to partial or complete human evaluation of the fitness of solutions generated from evolutionary search. This has been introduced where quantitative evaluation is difficult if not impossible to achieve. Examples of application include graphic arts and animation (Sims, 1991), food engineering (Herdy, 1997), and hazard icon design (Carnahan, 2004). Such applications rely upon a human-centred, subjective evaluation of the fitness of a particular design, image, taste, and so forth, as opposed to an evaluation developed from some analytic model.

Partial human interaction that complements quantitative machine-based solution evaluation is also evident—for instance, the user addition of new constraints in order to generate solutions that are fully satisfactory within an evolutionary nurse scheduling system (Inoue, Furuhashi, & Fujii, 1999). Another example is the introduction of new compounds as elite solutions into selected evolving generations of a biomolecular design process (Levine, Facello, & Hallstrom, 1997).

These examples utilise a major advantage of stochastic population-based search techniques—that is, their capabilities as powerful search and exploration algorithms that provide diverse, interesting, and potentially competitive solutions to a wide range of problems. Such solutions can provide information to the user which supports a better understanding of the problem domain whilst helping to identify best direction for future investigation (Parmee & Bonham, 1999), especially when operating within poorly defined decision-making environments. Extracted information supports development of the problem representation in an iterative, interactive evolutionary environment. Interactive evolutionary design systems (IEDSs) represent a human-centric approach (Parmee, 2002; Parmee, Watson, Cvetkovic, & Bonham, 2000) that generate and succinctly present information appertaining to complex relationships between the variables, objectives, and constraints that define a developing decision space.

In an attempt to categorise these various forms of IEC, it is possible to view complete human evaluation as explicit, whereas partial evaluation and interaction are less explicit, more subtle forms of human involvement. Completely implicit interaction occurs where users are unaware of their role.
in the evolution of a system (e.g., the Web-based tutorials of Semet, Lutton, Biojout, Jamont, & Collet, 2003). A simple implicit/explicit spectrum of interactive evolutionary approaches can thus be developed (Parmee & Abraham, 2005).

The following sections concentrate on the manner in which evolutionary search and exploration can generate high-quality information from complex design and decision-making environments. Such information can be utilised interactively to:

- support the identification of high-performance solutions where qualitative as well as quantitative objectives play a major role; and
- modify and refine design problem representation.

The first example relates to aesthetic judgement of EC-generated designs and is closer to the more traditional explicit interaction where user subjective evaluation is evident. However, this subjective evaluation complements detailed, machine-based quantitative evaluation. The second is current IEDS research relating to problem definition and the iterative interactive improvement of machine-based design representations that sits further toward the implicit end of the spectrum.

INTEGRATING AESTHETICS VIA INTERACTIVE EVOLUTIONARY DESIGN PROCESSES

This example brings together agent-based machine learning, evolutionary computing, and subjective evaluation in search for aesthetically pleasing, structurally feasible designs during the conceptual design process. Although significant theoretical work is evident with respect to the inclusion of aesthetics in computer-based design, application-based research has received less attention (Moore, Miles, & Evans, 1996a, 1996b; Saunders, 2001).

Figure 1 illustrates the main components of the system and the manner in which they interact. The user defines initial design requirements and aesthetically evaluates the designs generated by the Evolutionary Search, Exploration, and Optimisation System (ESEO) during the initial generations. The agents have multiple tasks, which include the creation of the initial population based on design requirements, the monitoring of designs for feasibility during the ESEO processes, and evaluation of machine-based aesthetic criteria. The ESEO identifies design solutions that can be considered high performance in terms of Structural Feasibility and Stability, Materials Cost, and Rule-Based Aesthetics.

*Figure 1. The user-centric system*
The project is initially considering three test domains: bridges; liquid containers such as wine glasses, chemical tanks, and so forth; and street furniture in the form of bench type structures. The following example concentrates on bridge design. Research involves the ACDDM Lab at Bristol UWE and the Institute of Machines and Structures at Cardiff University. The early stages have concentrated on development of highly flexible and robust representations of simple bridge structures and the identification of optimal solutions via basic evolutionary algorithms.

**Representation**

Problem representation affects search efficiency and evolutionary performance (Rosenman, 1997; Goldberg, 1989). Genetic algorithms (GAs), evolutionary strategies (ESs), and evolutionary programming (EP) representations are generally based on binary or real number parameter strings, but many alternatives are also available. Component-based and hierarchical representations support flexibility and robustness in terms of accommodating complex design entities with many related sub-systems/components (e.g., Bentley, 2000; Cramer, 1985; Rosenman, 1996).

It is also necessary to consider which stochastic search process would best suit a chosen representation in terms of efficient negotiation of the design space. As a high degree of solution search and exploration is required, a population-based approach would seem most appropriate. GAs, EPs, and ESs offer high utility with differing operators. GAs use crossover as the main operator (Goldberg, 1989). ESs are similar to real parameter GAs without crossover, although crossover-like operators have been introduced (Deb, 2001). EPs (Fogel, 1988) is a purely mutation-based evolutionary algorithm and therefore perhaps represents the simplest of the three, as the selected representation does not need to support crossover between differing variable strings. GAs, however, require a representation that is robust enough to handle repeated crossovers whilst ensuring offspring feasibility. ESs could be considered to lie somewhere between these two. EPs have been selected in this work primarily to overcome feasibility maintenance problems relating to the component-based representation and crossover.

The initial goal was a flexible enough representation to model all possible designs which is robust enough to be manipulated by design search and exploration processes (Machwe, Parmee, & Miles, 2005). A collection-based object-oriented representation has been developed. A single population member (chromosome) is a collection of primitive elements that represent a design. Different elements with different design properties can be included in the set of possible design primitives. The evaluation of fitness of the structure and checking the structural integrity utilises secondary properties of the particular primitive element type. Figure 2 further clarifies the idea of using an object-based representation. A simple bridge design is basically divided into

---

**Figure 2. Details of the object-based representation**
two separate collections. These are the *span element* collection containing elements which form the span of the bridge and the *support element* collection containing elements which form the support of the bridge. In the case of a simple, end-supported span, the support element collection will be empty.

The rectangular elements can either be part of a supporting element collection or a span element collection. Simple beam or angled beam span bridges with and without support require only two basic types of elements—the angled section element (to be used as a span element only) and a simple rectangle element, which can be used as both a spanning and supporting element. All elements have an assumed constant. The object-based representation can take advantage of the principles of object-oriented programming such as inheritance—that is, if we want to add a new kind of element, say curved span section, we could easily do so by extending the basic properties of element and adding the extra properties required for a curved section.

During initial design activity, it may not be desirable or possible to strictly parameterise a design. Designs make the transition from abstract concepts to well-defined specifications. The above object-based representation can represent designs at all levels, including at the abstract level by using high-level objects/elements and at the well-defined level as a set of specifications. Functionality can be modified based upon changing requirements. Thus objects offer a straightforward way to cover design activity at all levels. To produce the same in a simple string-based chromosome would require additional checks to ensure consistency of the chromosome is not violated once a bit or real number is added or removed, or its functionality modified. The overall system would be overly complex and difficult to manipulate.

**Mutation**

Let us assume the chromosome to be mutated represents the design in Figure 3.

This is a basic beam bridge simply supported at either end with a single intermediate support (B3) and two span elements (B1, B2). L is the span and H is the maximum height of the bridge (see Figure 4).

The mutation is rule based, and a rule is selected randomly. There are separate rules for the two arrays of elements. The supports can only move left or right. Their height is based upon the
thickness of the spanning element they support. Hence there are only four rules for supports: two rules for left and right movement, and two for increasing and decreasing width. The depth of each span element can vary, but they must have a level upper surface and must be continuous with no overlap or space between them. Thus for a simple element in a span, there are just two rules, namely to increase or decrease the span depth. Now, for example, if the selected rule for support (B3) is to move it left by a constant distance (say two units) and for span to decrease thickness of B2 support by constant units (say two units again), then the B3 object in the support array will have its X value attribute decreased by two units and the B2 object in the span array will have its height value attribute decreased by two units. The height attribute of the support will be automatically adjusted to make it continuous and remove any overlap at its new position. The mutated design is shown in Figure 5.

**Introduction of Agency**

Initial testing of the representation involved freeform assembly and evolution of simple bridge structures using GA, EP, and agent-based approaches. Evolving feasible structures proved a difficult and lengthy process, whereas rule-based agent assembly was straightforward and rapid. Thus a combined approach has been developed where agents create the initial population of structures and provide a continuous ‘repair’ capability, and an evolutionary system performs search, exploration, and optimisation across the space of possible structures.

The construction and repair agents (CARAs) have the task of assembling structures with varying size and shape regarding span and supports. Agents are given specifications relating to restrictions on placement of supports and types of span section. Thus CARAs are informed of the design environment and will create initial population designs within it. The evolutionary process then takes care of the SEO process, with the CARAs keeping a check on potentially disruptive processes and repairing the structure, where necessary, to ensure feasibility.

**Simple Structural Criteria**

The CARAs can currently create three kinds of bridges:

- simple beam bridges without supports (Type 1a);
- simple beam bridges with supports (Type 1b); and
- simple beam bridges with sloping span sections and supports (Type 2).

An initial population can consist of a mixture of these designs which are then evaluated against structural and aesthetic criteria. In terms of structure, solutions are assessed via simple length/depth ratios and minimising material. Column design is subjected to buckling criteria.
Fitness Evaluation for Type 1A

Type 1A is treated as a simple beam deflection problem under uniform distributed loading. A simple heuristic relating to an ideal length-to-height ratio for a span element of 20:1 is utilised—that is, the closer a span section is to the ideal ratio \( R \), the better its fitness.

\[ F_i = R - \left( \frac{L_i}{H_i} \right) \]  
\[ \text{Stability} = \frac{1}{(1 + \sum f_i)} \] (1)

To ensure the overall integrity of the structure, the above equations are used. It is evident that the closer the dimensions of the span elements are to the ideal ratio \( R \), the lower the value of \( F_i \) will be. At the minimum, all \( F_i \)'s are equal to zero and thus stability is equal to one.

Fitness Evaluation for Type 1b and Type 2

In Type 1b the buckling in the columns due to the weight of the loaded beam is also considered using:

\[ P' = \frac{\pi^2 EI}{H^2} \] (3)

where: \( P' \) = maximum possible load, \( E \) = modulus of elasticity, \( I \) = moment of inertia, and \( H \) = column height. The load on column is determined from the length of the beam between the end supports, calculating the loading and distributing this loading across intermediate supports. A column satisfying the buckling criteria can either increase or decrease in thickness. Otherwise it can only increase in thickness.

Example

A basic EP approach is used with a population size of 100 solutions. Tournament selection is utilised with a tournament size of 10, and the system is run for 100 generations. A few members from the initial population are shown in Figure 6.

The initial population consists of three different kinds of designs: a simply supported span, a span with columns, and an angled span bridge. After 100 generations the optimal designs shown in Figure 7 are achieved. The angled span bridges turn out to be most efficient in terms of structural criteria (i.e., stability and material usage). The
other two design types have been evolved out of the population.

Aesthetics and User Evaluation

It is very difficult to integrate subjective aesthetic criteria with machine-based design unless user interaction plays a significant role. Moore et al. (1996a, 1996b) have published significant work in this area which has contributed to a New South Wales RTA set of guidelines on bridge aesthetics (Road and Traffic Authority, 2004). However, aesthetics evaluation can only be partially quantified by generic guidelines and rules. While aesthetically pleasing shapes can be explicitly specified to some extent, complete aesthetic evaluation must also involve the designer—that is, evaluation must involve both rule-based and subjective factors. In the present system the following aesthetics have been hard coded:

- symmetry of support placement (A1),
- slenderness ratio (A2),
- uniformity in thickness of supports (A3), and
- uniformity in thickness of span sections (A4).

Many other such rules are currently being evaluated and included in the work as the design representation is developed to support detailed aesthetic evaluation. Each aesthetic rule is evaluated by a separate aesthetic agent. The rule-based aesthetic fitness is calculated as:

\[ \text{Aesthetic Fitness} = \sum_{i=1}^{4} w_i A_i \]  

(4)

where \( w_i \) are weights for each of the aesthetic rules (\( A_i = \text{A1 to A4} \)) which can also be modified online.

In addition, user-assigned aesthetic fitness (Ufit) is the fitness given to a design directly by the user on a scale of 0 to 10 (10 being the best). The user can also mark solutions for preservation into the next generation. Overall user evaluation operates thus:

1. User stipulates the frequency of user interaction (e.g., once every 10 generations).
2. User aesthetically evaluates a preset number of population members from the initial population (usually the top 10 members, i.e., those with highest fitness regarding stability, material usage, and explicitly defined aesthetic criteria).
3. The EP system runs.
4. Population members are aesthetically evaluated by the user every \( n \) generations.
5. Repeat steps 3 and 4 until user terminates the evolutionary process.

The overall fitness function now includes aesthetic fitness and user-assigned aesthetic fitness. Furthermore, weights have been added (w1 to w4) to each of the objectives which the user can modify online to influence evolutionary direction (See Box 1).

Figure 8 shows aesthetically pleasing cross-sections after 30 generations with user evaluation every 10 generations. The aesthetic objectives (A1 to A4) are clearly reflected in them. The span elements are of the same size. The supports are of nearly uniform thickness, and their placement is also symmetric.

Box 1.

\[ \text{Fitness} = (w1 \times \text{Stability}) + (\frac{w2}{\text{Material Usage}}) + (w3 \times \text{Aesthetic Fitness}) + (w4 \times \text{Ufit}) \]  

(5)
Figure 8. Aesthetically pleasing cross-sections

![Cross-sections](image)

Furthermore, due to user interaction, the optimised shapes are not limited to angled sections but take on a variety of different aesthetically pleasing shapes that not only satisfy the explicitly defined aesthetic guidelines (A1 to A4) but also the implicit aesthetics of the user (Ufit).

### Incorporating Learning

Current work involves the introduction of supervised learning taking user evaluation into account. Since there is a natural classification in the designs (i.e., angled spans, supported beams, and unsupported beams), learning is attempted at two levels. The first level determines user preference for one of the three types of bridge design. This is achieved by evaluating the relative difference between user-assigned fitness (or rank) for each type of design. The second level assesses what kind of features the user finds pleasing in the different designs. Figures 7 and 8 indicate that for the angled spans, there are two features that catch the eye immediately. These are the peak of the bridge (that is the location of the rise point) and the thickness of the span sections. Such features are converted into fuzzy variables to create an aesthetic model of the particular bridge type. For the angled section the following fuzzy variables are used to specify the aesthetic model:

1. **Peak**: Left, Central, Right
2. **Difference in Span Thickness**: Left Thicker, Equal Thickness, Right Thicker
3. **Average Thickness**: Low, Medium, High
4. **Column Thickness**: Low, Medium, High
5. **User-Assigned Fitness (Ufit)**: Low, Medium, High

Similar models can be created for supported beam spans and unsupported beam spans. Based on this model a fuzzy rule generator has been implemented. Initial results have been encouraging. The intention is that as search progresses there will be a gradual lessening of the degree of user interaction allied with an increasing degree of autonomous machine-based solution evaluation involving both aesthetic and structural criteria. This addresses the problem of user fatigue.

### EVOLVING THE PROBLEM SPACE THROUGH INTERACTIVE EVOLUTIONARY PROCESSES

The second illustration of the utilisation of user-centric evolutionary computing relates to a more implicit form of interaction. This is aimed primarily at the extraction of high-quality information and the succinct presentation of such information to the designer/decision maker in such a manner that supports a better understanding of complex relationships between variables, multiple objectives, and constraints during conceptual design. This complements the initial IED concept by further attempting to meld experiential knowledge and intuition with powerful machine-based search, exploration, and information processing.

Machine-based problem representations support exploration through the evaluation of solutions against seemingly relevant criteria. Although initial representations may be relatively basic, and confidence in model output will be low, such representations can provide essential problem insight despite apparent shortfalls. Identified high-performance solutions based upon quantitative criteria followed by qualitative human evaluation may provide an indication of concept viability and model fidelity. An iterative user/machine-based process can commence where gradual improvements in understanding contribute to the development of better representations, a growing knowledge base, and the establishment of computational models that support
rigorous analysis—that is, a process emerges that supports the development of representation through knowledge discovery.

An initial variable parameter set may be selected with later addition or removal of variables as the sensitivity of the problem to various aspects becomes apparent. Constraints may be treated in the same way with the added option of softening them to allow exploration of non-feasible regions. Included objectives may change as significant payback becomes apparent through a reordering of objective preferences. Some non-conflicting objectives may merge, whilst difficulties relating to others may require serious re-thinking with regard to problem formulation. The initial design space is therefore a moving feast rich in information (Parmee, 2002).

The visualisation of variable and objective space from cluster-oriented genetic algorithm (COGA) output provides a variety of perspectives illustrating complex relationships (Parmee & Abraham, 2004). This information is further defined by data mining, processing, and visualisation techniques. The intention is to support implicit learning and reduce complexity by supporting the designer development of a quantitative and intuitional understanding of the problem. This leads to the iterative model development described above.

COGAs and the MiniCAPs Model

Cluster-oriented genetic algorithms provide a means to identify high-performance (HP) regions of complex conceptual design spaces and enable the extraction of information from such regions (Parmee, 1996). COGAs identify HP regions through the online adaptive filtering of solutions generated by a genetic algorithm. COGA can be utilised to generate design information relating to single and multi-objective domains (Parmee & Bonham, 1999). The technique has been well documented (see http://www.ad-comtech.co.uk/Parmee-Publications.htm for relevant papers).

The research utilises the BAE Systems’ MiniCAPs model, a simplified version of a suite of preliminary design models for the early stages of military aircraft airframe design and initially developed for research relating to the development of the IED concept. The model comprises nine continuous input variables and 12 continuous output parameters relating to criteria such as performance, wing geometry, propulsion, fuel capacity, structural integrity, and so forth. Input variables are:

1. Climb Mach Number (CLMN)
2. Cruise Height (CH)
3. Cruise Mach Number (CRMN)
4. Gross Wing Plan Area (GWP)
5. Wing Aspect Ratio (WAR)
6. Wing Taper Ratio (WTR)
7. Wing Lead Edge Sweep (WLES)
8. Wing T/C Ratio (WTCR)
9. By Pass Ratio (BPR)

Identifying High-Performance Regions Relating to Differing Objectives

Figures 9(a), (b), and (c) show HP regions comprising COGA-generated solutions relating to three of the 12 MiniCAPS objectives (Ferry Range—FR, Attained Turn Rate—ATR1, and Specific Excess Power—SEP1) projected onto a variable hyperplane relating to two of the nine variables utilised in the search process. This projection allows the designer to visualise the HP regions, identify their bounds, and subsequently reduce the variable ranges as described in previously referenced papers. These papers also introduce the projection of these differing objective HP regions onto the same variable hyperplane as shown in Figure 10 from which the degree of objective conflict immediately becomes apparent to the designer. The emergence of a mutually inclusive region of HP solutions relating to the ATR1 and FR objectives indicates...
Figure 9. COGA-generated high-performance regions relating to three differing objectives: (a) FR—Ferry Range; (b) ATR1—Attained Turn Rate; (c) SEP1—Specific Excess Power. All projected onto the GWPA (Gross Wing Plan Area)/WAR (Wing Aspect Ratio) variable hyperplane (N.B. Colour versions of figures can be found at: http://www.ad-comtech.co.uk/cogaplots.htm)

![Graph](image)

(a) 

(b) 

(c)

a low degree of conflict, whereas the HP region relating to SEP1 is remote (in variable space) to both the ATR1 and FR regions, indicating a higher degree of conflict.

There is much information contained in the HP regions relating to appropriate variable ranges for single objectives, degree of conflict between multiple objectives, and the emergence and definition of mutually inclusive (common) HP regions. This graphical representation provides an excellent spatial indication of the degree of objective conflict. However, searching through all possible two-dimensional variable hyperplanes to visualise such information is not a feasible approach. Recent research has resulted in single graphical representations that can present all variable and objective data whilst providing links to other visual perspectives. The parallel coordinate box plot (PCBP) representation shown in Figure 11 is one such graphic that provides a central repository containing much single and multiple-objective solution information.
Parallel Coordinate Box Plots

Parallel coordinate plots (Inselberg, 1985) appeared to offer potential in terms of providing a single graphic illustrating complex relationships between variable and objective space. Parallel coordinate representation displays each variable dimension vertically parallel to each other. Points corresponding to a solution's value of that variable can then be plotted on each vertical variable axis. It is thus possible to show the distribution of solutions in all variable dimensions and the correlation between different dimensions. The disadvantage of the technique when attempting to include multiple objectives is that the density of the information presented hinders perception (Parmee & Abraham, 2004, 2005). To overcome the ‘data density’ problem, three modifications to the standard parallel coordinate representation have been included:

1. additional vertical axes for each variable so that each objective can be represented,
2. an indication of the degree of HP region solution cover across each variable range, and
3. the introduction of box plots to indicate skewness of solutions across each variable range.

This PCBP provides a much clearer graphic (see Figure 11). The vertical axis of each variable is scaled between the minimum and maximum value of the variable in the HP region solutions of each objective. The length of the axis represents the normalised ranges of variable values present in an HP region. Where an HP solution set does not fully extend across the variable range, the axis is terminated by a whisker at the maximum or minimum value of the variable. The colour-coded box plots relate to each objective (i.e., SEP1, ATR1, and FR). The median is marked within the box, and the box extends between the lower and upper quartile values within the variable set. The PCBP clearly visualises the skewness of solution distribution relating to each objective in each variable dimension which provides an indication of the degree of conflict between objectives.

For instance, it is apparent that all three objective boxes overlap in the case of variables 1, 2, 3, 6, and 9. However, significant differences in the distribution of the boxes are evident in terms of at least one objective where variables 4, 5, 7, and 8 are concerned. Variables 4 and 5 are Gross Wing Plan Area and Wing Aspect Ratio. The conflict between SEP1 and FR/ATR1 evident in Figure 10 is strongly reflected in the HP solution distribution indicated by the whisker truncation of variable 4 in Figure 11 and in the box plots of that variable. In terms of variable 5, the whisker terminations relating to ATR1 and FR in Figure 11 reflect the extent of the solution distribution across their HP regions in Figure 10. The box plots also reflect the relative distribution of HP solutions of all objectives along that variable plane as illustrated in Figure 10. Figure 12 shows a projection of the ATR1 HP region onto the Cruise Height (variable 1) and Climb Mach No (variable 2) hyperplane. The relatively uniform distribution of HP solutions across the hyperplane is reflected in the appropriate variable plots of Figure 11.

The PCBP represents a single graphic from which the designer can perceive which variables are causing high degrees of objective conflict. To get an alternative, very clear perspective of these conflicts, any two of these variables can be selected to also view the relevant graphics similar to Figure 10. Further reinforcement can be obtained from the perspectives explored in the following section relating to projections upon objective space. Improved understanding can lead to developments of the computational representation and to appropriate setting of objective preferences.
Projection of COGA Output on to Objective Space

The HP region solutions for ATR1 and FR can be projected onto objective space as shown in Figure 13. A relationship between the HP region solutions and a Pareto-frontier emerges along the outer edge of the plot (Parmee & Abraham, 2004) despite the fact that the working principle of COGA is very different to that of evolutionary multi-objective algorithms (Deb, 2001), which tend to use a non-dominance approach.

For comparative purposes, Figure 14 illustrates the distribution of COGA output and SPEA-II (Zitzler et al., 2002) Pareto-front output in objective space. Using a standard multi-objective GA (MOGA), it is possible to obtain solutions lying along the Pareto-front, but difficult to explore the relationship between variable and the objective space. However, it is likely that the designer is also interested in solutions that lie around particular sections of the Pareto-front.
The COGA approach therefore provides a good visual indication of the degree of conflict between objectives, an opportunity to explore varying objective preferences and view their effect upon HP region bounds, and the ability to generate an approximate Pareto-front relating to the objectives under investigation plus solutions around the Pareto-front. This is in addition to the utility of COGA in single-objective space as described in previous referenced papers. All of this utility directly complements the original IEDS concept regarding information extraction, processing, and presentation.

SUMMARY AND CONCLUSION

The aesthetics work reveals a significant potential in terms of the development of systems that include criteria ranging from purely quantitative through to purely subjective. Ultimately the system will be required to give a comparative indication in terms of aesthetically pleasing design and likely cost whilst indicating structural feasibility.

The introduction of such an interactive process also poses many questions such as:

- How many designs from each population should be presented to the user?
- How should these be selected?
- How many evaluations can a user be expected to perform before becoming fatigued?

These questions have been repeatedly posed, but seldom successfully addressed within the interactive evolutionary computing (IEC) community. Our continuing research is addressing these issues and, in particular, the user-fatigue aspect.

The integration of user preference and user-varied objective weights supports the transfer of subjective evaluation from the user to a design/decision-making system. In order to address the third question above, a machine learning system is required which learns the preferences of the user during the interactive process. User fatigue is perhaps the major stumbling block in the development of successful systems. Much work is required to develop an appropriate, fully functioning, machine-learning sub-system. The fuzzy rule-based learning system is not ideal, and current work is investigating a case-based approach.

The developing system should be seen as a generic framework for the integration of user-evaluation with any preliminary design/decision-making domain. The CARA-EP representation concept should be portable across many problem domains. Any system must significantly decrease the load on the user as early as possible in the evolutionary process. A multi-agent-based learning environment is therefore under investigation that gradually replaces the subjective criteria evaluation.

It is apparent from previous research and the research presented here that COGA-generated data can provide visual representations in variable space of the degree of conflict between objectives and excellent spatial indications of the distribution of high-performance solution regions relating to
a number of objectives. It is also apparent that the COGA HP solution sets, when projected onto objective space, provide the designer with an opportunity to explore a wealth of HP solutions that offer varying degrees of objective compromise and a variety of design characteristics. The non-dominance sorting of these solutions also provides an approximate Pareto-frontier illustrating succinct available trade-offs. The direct mapping of solutions between objective and variable space facilitates an understanding of the relative utility of solutions in terms of preferred variable ranges and particular design characteristics.

The PCBP of Figure 11 offers a first point of call for the designer to get an overview of the varied information available from COGA output. The intention is that the COGA graphical perspectives will be available through simple menu/clicking operations from the central PCBP image. These differing perspectives are seen as essential aids to understanding overall complexities relating to the two dependant design spaces (variable vs. objective space).

There is a wealth of information available from COGA output relating to single objective solutions that is also inherent within the multi-objective output. Hence the utility of the approach should be assessed across both areas. The information available from single-objective HP regions has been fully discussed in previous referenced papers.

User-centric techniques described in the chapter and variations of them are also currently being applied in the conceptual design of submersible vehicles (Parmee & Abraham, 2005), pharmaceutical drug design and discovery (Sharma & Parmee, 2005), and conceptual software design (Simons & Parmee, 2004). Details of this associated work can be found on the ACDDM Web site at www.ad-comtech.co.uk/ACDDM_Group.htm.

ACKNOWLEDGMENT

The authors wish to acknowledge the contribution to the bridge aesthetics work of Professor John Miles of the Institute of Machines and Structures at the University of Cardiff.

REFERENCES


Chapter 2.3
Adaptation and Personalization of User Interface and Content

Christos K. Georgiadis
University of Macedonia, Thessaloniki, Greece

ABSTRACT

Adaptive services based on context-awareness are considered to be a precious benefit of mobile applications. Effective adaptations however, have to be based on critical context criteria. For example, presence and availability mechanisms enable the system to decide when the user is in a certain locale and whether the user is available to engage in certain actions. What is even more challenging is a personalization of the user interface to the interests and preferences of the individual user and the characteristics of the used end device. Multimedia personalization is concerned with the building of an adaptive multimedia system that can customize the representation of multimedia content to the needs of a user. Mobile multimedia personalization especially, is related with the particular features of mobile devices’ usage. In order to fully support customization processes, a personalization perspective is essential to classify the multimedia interface elements and to analyze their influence on the effectiveness of mobile applications.

INTRODUCTION

Limited resources of mobile computing infrastructure (cellular networks and end user devices) set strict requirements to the transmission and presentation of multimedia. These constraints elevate the importance of additional mechanisms, capable of handling economically and efficiently the multimedia content. Flexible techniques are needed to model multimedia data adaptively for multiple heterogeneous networks and devices with varying capabilities. “Context” conditions (the implicit information about the environment, situation and surrounding of a particular communication) are of great importance.

Adaptive services based on context-awareness are indeed a precious benefit of mobile applications: in order to improve their provided service, mobile applications can actually take advantage of the context to adjust their behaviors. An effective adaptation has to be based on certain context criteria: presence and availability mechanisms enable the system to decide when the user is in a certain locale and whether the user is available
to engage in certain actions. Hence, mobile applications aim to adapt the multimedia content to the different end user devices.

However, typically each and every person receives the same information under the same context conditions. What is even more challenging is a personalization of the user interface (UI) to the interests and preferences of the individual user and the characteristics of the user end device. The goal of mobile applications is to increasingly make their service offerings more personalized toward their users. Personalization has the ability to adapt (customize) resources (products, information, or services) to better fit the needs of each user. Personalization in mobile applications enables advanced customized services such as alerts, targeted advertising, games, and improved, push-based mobile messaging. In particular, multimedia personalization is concerned with the building of an adaptive multimedia system that can customize the representation of multimedia content to the needs of a user.

Multimedia personalization enlarges the application’s complexity since every individual’s options have to be considered and implemented. It results in a massive amount of variant possibilities: target groups, output formats, mobile end devices, languages, locations, etc. Thus, manual selection and composition of multimedia content is not practical. A “personalization engine” is needed to dynamically create the context-dependent personalized multimedia content. General solution approaches concerning the personalization engine, include personalization by transformation (using XML-based transformations to produce personalized multimedia documents), adaptive multimedia documents (using SMIL-like presentation defined alternatives), personalization by constraints (optimization problem—constraint solving), personalization by algebraic operators (algebra to select media elements and merge them into a coherent multimedia presentation), or broader software engineering approaches.

Mobile multimedia (M3) personalization especially, is related with the particular features of mobile devices’ usage. Because of their mobility and omnipresence, mobile devices have two characteristics worth noticing. First, users have limited attention as they operate their mobile devices (this is because they usually are concerned at the same time in other tasks, e.g., car driving). Second, users tend to treat their mobile devices in a quite personal way, seeking for personal services and personalized content. The preferences of users are therefore noticeably affected. In many cases, they favor content and services which do not require transmitting large quantities of information. Thus, low-intensity content (e.g., ring tones, weather reports, and screen icons) proved to be very popular. This is not only because of the low availability of mobile devices’ resources which complicates the processing of large volumes of information. Users demand further individually customized content on the mobile Internet because its personalization level is higher than that of the fixed Internet.

Detailed issues concerning M3 personalization can be described, analyzing UI design issues. Existing mobile applications offer a reasonably easy, browser-based interface to help user access available information or services. In order to support adaptation and personalization mechanisms they should be also as far as possible concentrated on the individual prerequisites of the human in contact with it. In this chapter, after the presentation of background topics we discuss critical issues of the mobile setting (characteristics of mobile applications and mobility dimensions in user interactions) that influence adaptation and personalization technologies. Then, as an application case, we focus on m-commerce applications and customer interfaces. All current research studies tend to acknowledge that the design rules of wired Internet applications are only partially useful. They should not be directly adopted in mobile computing area, because of the considerably different user requirements and device constraints. On the other hand, experi-
Adaptation and Personalization of User Interface and Content

ence gained from the fixed Internet formulated as the well-accepted 7C framework, is always welcomed. Hence, we classify the multimedia interface elements and we analyze their influence on m-commerce site’s effectiveness from a personalization perspective.

BACKGROUND

Adaptation Objectives

The diversity of end device and network capabilities in mobile applications along with the known multimedia challenges (namely, the efficient management of size, time, and semantics parameters of multimedia), demand media content and service to be flexible modeled for providing easy-to-use and fast multimedia information. Multimedia adaptation is being researched to merge the creation of the services so that only one service is needed to cover the heterogeneous environments (Forstadius, Ala-Kurikka, Koivisto, & Sauvola, 2001). Even though adaptation effects could be realized in a variety of ways, the major multimedia adaptation technologies are adaptive content selection, and adaptive presentation. Examples of adaptation include “down-scaling” the multimedia objects and changing the style of multimedia presentation according to user’s context conditions.

In general, adaptive hypermedia and adaptive Web systems belong to the class of user-adaptive systems. A user model—the explicit representation of all relevant aspects of a user’s preferences, intentions, etc.—forms the foundation of all adaptive systems (Bauer, 2004). The user model is used to provide an adaptation effect, which is tailoring interaction to different users. The first two generations (pre-Web and Web) of adaptive systems explored mainly adaptive content selection and adaptive recommendation based on modeling user interests. Nowadays, the third (mobile) generation extends the basis of the adaptation by adding models of context (location, time, bandwidth, computing platform, etc.) to the classic user models and explores the use of known adaptation technologies to adapt to both an individual user and the context of the user’s work (Brusilovsky & Maybury, 2002).

Personalization Objectives and Mechanisms

Personalization is a special kind of adaptation of the UI which focuses on making a Web application more receptive to the unique and individual needs of each user (Cingil, Dogac, & Azgin, 2000). Personalization mechanisms presuppose two phases. First, the accumulation of user information, in order to build up a profile that illustrates a set of descriptors essential to administrators (e.g., visitor’s interest, navigation paths, entitlements and roles in an organization, purchases, etc.). The second phase is the analysis of user information to recommend actions specific to the user.

To develop the best recommendation, rule-based practices (allowing administrators to specify principles for their applications to thrust personalization) are usually combined with filtering algorithms which analyze user profiles (Pierakos, Paliouras, Papatheodorou, & Spyropoulos, 2003). Simple filtering techniques are based on predefined groups of users, classifying their accounts by age groups, asset value etc. Content-based filtering can be seen as locating objects comparable to those a user was fond of in the past. Finally, collaborative filtering builds up recommendations by discovering users with similar preferences.

CONTENT ADAPTATION AND PERSONALIZED USER INTERFACE

Analyzing Mobile Setting

The characteristics of the mobile Internet applications can be appreciated from three different
Adaptation and Personalization of User Interface and Content

viewpoints: system, environment, and user (Chae & Kim, 2003). From the system’s viewpoint, mobile applications present disadvantages, because they provide a lower level of available system resources. Mobile devices, especially cellular phones, have lower multimedia processing capabilities, inconvenient input/output facilities (smaller screens/keyboards), and lower network connection speeds than desktop computers. However, from the environmental viewpoint there is an uncontested benefit: they enable users to access mobile Internet content anywhere and anytime. The term “instant connectivity” is used for mobile browsing to describe actually the fact that it is possible to do it at the moment of need.

User’s perspective characteristics must be regarded rather differently, because they are to a certain degree consequences of the system and of the environment. In addition, the multi-dimensional concept of “mobility” influences on them in many ways. Mobile users perform their tasks in terms of place, time and context. Different terms are used by the research community to describe user’s mobile setting, and their interactions within it, but these converge at the ones described below (Kakihara & Sorensen, 2001; Lee & Benbasat, 2004):

Figure 1. Analyzing mobile setting
Spatial mobility denotes mainly the most immediate dimension of mobility, the extensive geographical movement of users. As users carry their mobile devices anywhere they go, spatiality includes the mobility of both the user and the device.

Temporal mobility refers to the ability of users for mobile browsing while engaged in a peripheral task.

Contextual mobility signifies the character of the dynamic conditions in which users employ mobile devices. Users’ actions are intrinsically situated in a particular context that frames and it is framed by the performance of their actions recursively.

Because of their mobility (and in correspondence with its dimensions), we distinguish three attributes regarding mobile device usage:

1. Users have a tendency to treat their mobile device in a quite personal and emotional way (Chae & Kim, 2003). They prefer to access more personalized services when they are involved in mobile browsing. Spatial mobility must be considered as the major reason behind this behaviour, which is quite normal considering user’s perspective: the mobile phone is a portable, ubiquitous and exposed on everybody’s view gadget, able to signify user’s aesthetic preferences and personality.

2. Users have limited attention as they manage their mobile devices (Lee & Benbasat, 2004). This is because they usually are involved at the same time in other tasks (e.g., walking). Temporal mobility is the reason of this phenomenon.

3. Users manage their mobile devices in broadly mixed environments that are relatively unsteady from one moment to the next. Contextual mobility requires context-sensitivity on mobile device operations. So, mobile device is able to detect the user’s setting (such as location and resources nearby) and subsequently to propose this information to the mobile application. In this way, mobile device practically may offer task-relevant services and information.

Application Case: User Interfaces in M-Commerce Applications

Mobile Commerce Applications

The mobile sector is creating exciting new opportunities for content and applications developers. The use of wireless technologies extends the nature and scope of traditional e-commerce by providing the additional aspects of mobility (of participation) and portability (of technology) (Elliot & Phillips, 2004). One of the most rapidly spreading applications within the m-commerce world is the mobile Internet: the wireless access to the contents of the Internet using portable devices, such as mobile phones. Undoubtedly, delivering personalized information is a critical factor concerning the effectiveness of an m-commerce application: the organization knows how to treat each visitor on an individual basis and emulate a traditional face-to-face transaction. Thus, has the ability to treat visitors based on their personal qualities and on prior history with its site. M-commerce applications support mechanisms to learn more about visitor (customer) desires, to recognize future trends or expectations and hopefully to amplify customer “loyalty” to the provided services.

Personalized Multimedia in Interfaces of M-Commerce Applications

The goal of adaptive personalization is to increase the usage and acceptance of mobile access through content that is easily accessible and personally relevant (Billsus, Brunk, Evans, Gladish, & Pazzani, 2002). The importance of interface design has been commonly acknowledged, especially
Adaptation and Personalization of User Interface and Content

Lee and Benbasat (2004) describe in detail the influence of mobile Internet environment to the 7C framework for customer interfaces. This framework studies interface and content issues based on the following design elements: customization (site’s ability to be personalized), content (what a site delivers), context/presentation (how is presented), connection (the degree of formal linkage from one site to others), communication (the type of dialogues between sites and their users), community (the interaction between users), and commerce (interface elements that support the various business transactions) (Rayport & Jaworski, 2001). A generic personalized perspective is presented in (Pierrakos et al., 2003) with a comprehensive classification scheme for Web personalization systems. Based on all these works, we focus on multimedia design issues concerning personalized UIs for m-commerce applications. We present a reconsideration of the 7C framework from an M3 customization aspect, in which we distinguish the following mobile multimedia adaptation/personalization categories:

**M3 content** is the main category. It contains the parts of the 7C’s “content” and “commerce” design elements, which deal with the choice of media. “Multimedia mix” is the term that is used in 7C framework regarding exclusively the ‘content’ element. However, in our approach regarding mobile devices adoption: interfaces characteristics had been identified as one of the two broad factors (along with network capabilities), affecting the implementation and acceptance of mobile phones emerged (Sarker & Wells, 2003). Devices adoption is a critical aspect for the future of m-commerce, because without widespread proliferation of mobile devices, m-commerce can not fulfill its potential.
multimedia elements regarding shopping carts, delivery options etc. are also belong here because they share a lot of commons concerning adaptation and personalization. It is commonly accepted that large and high visual fidelity images, audio effects, and motion on interfaces are multimedia effects which might lead to a higher probability of affecting users’ decisions in e-commerce environments (Lee & Benbasat, 2003). However, in m-commerce setting things are different because we can not assume that the underlying communication system is capable of delivering an optimum quality of service (QoS). The bandwidth on offer and the capabilities of devices are setting limitations. Therefore, a central issue to the acceptance of multimedia in m-commerce interfaces is the one of quality. The longer the response delay, the less inclined will the user be to visit that specific m-commerce site, resulting in lost revenue. Obviously, an end-to-end QoS over a variety of heterogeneous network domains and devices is not easily assured, but this is where adaptation principle steps in.

Dynamic content adaptation of the media quality to the level admitted by the network is a promising approach (Kosch, 2004). Content adaptation can be accomplished by modifying the quality of a media object (resolution and its play rate); so, it can be delivered over the network with the available bandwidth and then it can be presented at the end device (satisfying its access and user constraints).

An essential issue for effective content adaptation is the perceptual quality of multimedia. Quality of perception (QoP) is a measure which includes not only a user’s satisfaction with multimedia clips, but also his ability to perceive, analyze, and synthesize their informational content. When a “personalization engine” is called out to adapt multimedia content, the perceptual impact of QoS can be extremely valuable, and it can be summarized by the following points (Ghinea & Angelides, 2004):

- Missing a small number of media units will not be negatively perceived, given that too many such units are not missed consecutively and that this incident is infrequent
- Media streams could flow in and out of synchronization without substantial human displeasure
- Video rate variations are tolerated much better than rate variations in audio
- Audio loss of human speech is tolerated quite well
- Reducing the frame rate does not proportionally reduce the user’s understanding (user has more time to view a frame before changes)
- Users have difficulty absorbing audio, textual and visual information concurrently, as they tend to focus on one of these media at any one moment (although they may switch between the different media)
- Highly dynamic scenes have a negative impact on user understanding and information assimilation

Another important issue regarding M3 content adaptation (both for quality and for the selection of media items), is the usage patterns for the mobile Internet. Users purchase more low-risk products (e.g., books) than high-risk ones, because they can not pay full attention to their interactions with mobile devices. Also, users tend to subscribe to content with low information intensity more than to content with high information intensity (e.g., education), because mobile devices have inferior visual displays.

Device’s constraints and personalization requirements emphasize the need for additional effective content adaptation methods. Personalization mechanisms allow customers to feel sufficiently informed about products and services they are interested in, despite the limited multimedia information delivered by a restricted display device. They can be considered as filters which reject the delivery of multimedia content
Adaptation and Personalization of User Interface and Content

that users don’t appreciate. More and more, mobile applications exploit positioning information like GPS to guide the user on certain circumstances providing orientation and navigation multimedia information, such as location-sensitive maps. To facilitate personalized adaptation, multimedia content is desirable to include personalization and user profile management information (in the form of media descriptors) (Kosch, 2004). In this way, adaptive systems can utilize information from the context (or user model) in use. Especially personalized UIs are able to exercise all kinds of personalization mechanisms (rule-based practices and simple, content-based or collaborative filtering), to locate or predict a particular user’s opinion on multimedia items.

**M3 presentation** is also an important adaptation/personalization category. It contains all parts from 7C’s “context,” “commerce,” and “connection” design elements related to multimedia presentation. M3 presentation refers to the following aspects:

- The aesthetic nature of multimedia in interfaces (i.e., the visual and audio characteristics such as color schemes, screen icons, ring melodies, etc.). These multimedia UI elements are certainly used by mobile users in order to make their phones more personal
- The operational nature of multimedia in interfaces, including internal/external link issues and navigation tools (in what ways the moving throughout the application is supported). An important issue here deals with the limited attention of users when interacting with their mobile devices. So, minimal attention interface elements, able to minimize the amount of user attention required to operate a device are welcomed. For example, utilizing audio feedback in order to supplement users’ limited visual attention is considered in general a desirable approach in mobile setting (Kristoffersen & Ljungberg, 1999). There is also an additional point to take into consideration regarding M3 presentation adaptation and personalization: how to overcome the limitations due to the lack of screen space. Certainly, visual representations of objects, mostly through graphic icons, are easier to manipulate and retain than textual representations. But, small screens need not set aside a large portion of their space for infrequently used widgets. In this context, potential adaptations can be made by substituting visual elements with non-speech audio cues (Walker & Brewster, 1999), or by using semi-transparent screen-buttons, that overlap with the main body of content in order to make the most of a small screen (Kamba, Elson, Harpold, Stamper, & Sukaviriya, 1996).

All users are not having the same context conditions and preferences. Personalization mechanisms are used for both the aesthetic and the operational nature of multimedia in interfaces. Obviously, multimedia personalization engine must be able to provide context-sensitive personalized multimedia presentation. Hence, when a distracting user setting is acknowledged, the adapted multimedia presentations on the interface should call for only minimal attention in order to complete successfully critical transaction steps. Moreover, context-awareness of mobile devices may influence M3 presentation adaptation/personalization regarding connection issues. Indeed, the recommendation of a particular external link among a set of similar ones may depend not only from its content, but also from its availability and efficiency under the specific conditions of user’s setting.

**M3 communication** contains all parts from 7C’s “communication” and “community” design elements related to multimedia. In our approach, they belong to the same adaptation/personalization category because they deal with multimedia enriched communication and interaction services.
Mobile devices are inherently communication devices. Location and positioning mechanisms provide precise location information, enabling them to better interact with applications to deliver greatly targeted multimedia communication services. The perceptual quality of multimedia and relative previously discussed issues, are also important factors for effective multimedia communication adaptation.

With M3 communication personalization, m-commerce administrators are able to make use of information about users’ mobile setting to catch the right type of multimedia communication for the right moment (taken into account also the preferences of each user about the most wanted type of communication between him or her and the site). In addition, supporting adaptive (interactive or non-interactive) multimedia communication between users enables opinion exchange about current transactions and network accesses. Undoubtedly, such functionality may provide useful information for collaborative filtering techniques, resulting in more successful personalized sites.

**FUTURE TRENDS**

Providing adaptation and personalization affects system performance, and this is an open research issue. A basic approach to improve performance is to cache embedded multimedia files. However, when personalized multimedia elements are used extensively, multimedia caching can not maximize performance. The trend is therefore to provide personalization capabilities when server-usage is light and disallow such capabilities at periods of high request. Alternatively, users can have a personalized experience, even at times of high system load, if they pay for the privilege (Ghinea & Angelides, 2004). In any case, the design of a flexible context (or user) model, capable of understanding the characteristics of mobile setting in order to facilitate multimedia adaptation and personalization processes, it appears as an interesting research opportunity.

In a multi-layered wireless Web site, more sophisticated adaptation and personalization mechanisms are introduced as we get closer to the database layer. From that point of view, multimedia database management system (MMDBMS) emerging technology may support significantly the (mobile) multimedia content adaptation process. Existing multimedia data models in MMDBMSs are able to partially satisfy the requirements of multimedia content adaptation because contain only the basic information about the delivery of data (e.g., frame rate, compression method, etc.). More sophisticated characteristics such as the quality adaptation capabilities of the streams are not included. This information would be of interest to the end user. Consequently, a lot of research deals with extending the functionalities of current MMDBMSs by constructing a common framework for both the quality adaptation capabilities of multimedia and for the modeling/querying of multimedia in a multimedia database (Dunkley, 2003; Kosch, 2004).

**CONCLUSION**

The advances in network technology, together with novel communication protocols and the considerably enhanced throughput bandwidths of networks, attracted more and more consumers to load or stream multimedia data to their mobile devices. In addition, given the limited display space, the use of multimedia is recommended so that display space can be conserved. However, mobile setting’s limitations regarding multimedia are serious. In fact, enhancing the mobile browsing user experience with multimedia is feasible only if perceptual and contextual considerations are employed.

The major conclusion of previously presented issues is that efficient delivery, presentation and transmission of multimedia has to rely on context-
sensitive mechanisms, in order to be able to adapt multimedia to the limitations and needs of the environment at hand, and even more to personalize multimedia to individual user’s preferences.

REFERENCES


KEY TERMS

Content Adaptation: The alteration of the multimedia content to an alternative form to meet current usage and resource constraints.

MMDBMS: Multimedia database management system is a DBMS able to handle diverse kinds of multimedia and to provide sophisticated mechanisms for querying, processing, retrieving, inserting, deleting, and updating multimedia. Multimedia database storage and content-based search is supported in a standardized way.

Personalization: The automatic adjustment of information content, structure and presentation tailored to an individual user.

QoS: Quality of service notes the idea that transmission quality and service availability can be measured, improved, and, to some extent, guaranteed in advance. QoS is of particular concern for the continuous transmission of multimedia information and declares the ability of a network to deliver traffic with minimum delay and maximum availability.

Streaming: Breaking multimedia data into packets with sizes suitable for transmission between the servers and clients, in order to allow user to start enjoying the multimedia without waiting to the end of transmission.

UI: (Graphical) user interface is the part of the computer system which is exposed to users. They interact with it using menus, icons, mouse clicks, keystrokes and similar capabilities.

This work was previously published in Handbook of Research on Mobile Multimedia, edited by I. Ibrahim, pp. 266-277, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 2.4
User Interaction and Interface Design with UML

Jesus M. Almendros-Jimenez
Universidad de Almería, Spain

Luis Iribarne
Universidad de Almería, Spain

ABSTRACT
This chapter will show you how to use and specialise UML diagrams for describing the user interfaces of a software system. In order to accomplish the description of user interfaces, the proposed technique considers three specialised UML diagrams called user-interaction, user-interface, and GUI-class diagrams, which will be built following a model-driven development (MDD) perspective. These diagrams can be seen as the UML-based UI models of the system. In addition, this chapter is concerned with code-generation to implement the user interfaces of the system by using GUI-class diagrams and user-interaction diagrams. A case study of an Internet book shopping system is introduced in this chapter to proof and illustrate the proposed user interaction and interface design technique.

INTRODUCTION
The emergence of the unified modelling language (UML) (OMG, 2005) as an industry standard for modelling systems has encouraged the use of automated software tools that facilitate the development process from analysis through coding. The user interface (UI), as a significant part of most applications, should also be modelled using UML. UML diagrams could be used to model user interfaces, and automatic CASE tools could help to generate code for user interfaces from UML designs. In general terms, visual modelling allows the developers to visualize source code in a graphical form: graphical abstractions, such as flow charts to depict algorithmic control flows and structure charts or simple block diagrams with boxes representing functions and subprograms, and so on. UML provides system
User Interaction and Interface Design with UML

architects with a visual language for specifying, constructing, and documenting the artefacts of software systems. In particular, user interfaces should be visually modelled in order to describe the behaviour of the window system in response to user interactions.

This chapter is firstly devoted to show how to use and specialise UML diagrams in order to describe the user interface and user interactions of a software system, following a particular model-driven development (MDD) perspective. Model-driven development involves creating models through a methodological process that begins with requirements and looks into a high-level architectural design. Model-driven development facilitates and improves the software analysis and design and code generation facilities from models prevent the loss of substantial information during the transition of a model to its implementation.

In our MDD perspective, we consider the following steps for user interface design and modelling:

1. Firstly, we use a UML use case diagram for extracting the main user interfaces.
2. Secondly, we describe each use case by means of a special kind of UML activity diagrams, called user-interaction diagrams, whose states represent data output actions and transitions represent data input events. This perspective allows the designer to model the user interaction (i.e., input-output interaction) in each main user interface.
3. Thirdly, each input and output interaction of the user-interaction diagrams allows the designer to extract GUI components used in each user interface. Therefore, we can obtain a new and specialized version of the use case diagram representing the user interface design, and a class diagram for GUI components: user-interface and GUI-class diagrams, respectively.
4. The user-interaction, user-interface, and GUI-class diagrams can be seen as the UML-based user interface models of the system.

This chapter will also deal with code generation techniques. In our MDD perspective, the UML-based user interface models can be used for generating executable code with the following advantages:

1. Rapid prototyping of the developed software: Software modellers would find it useful to quickly generate user interfaces from high-level descriptions of the system.
2. Model validation and refinement: Prototyping can detect fails in design and refinement and validation of model by testing user interfaces and user requirements.
3. Model-based code generation: Generated code would fit with developed models.
4. Starting point for implementers: Prototypes can be refined until final implementation.

BACKGROUND

In the literature there are some works dealing with the problem of user interfaces in UML.

Use Cases and UI Design

Some of these works (CommentEdition, 2000; Constantine & Lockwood, 2001; Nunes & Falcao, 2001; Paterno, 2001) are focused on the utilisation of UML use case diagrams as a “starting point” of the user interface design, or even as a “high-level description” of the structure of the user interface. However, there are some considerations about the use case diagram style. Following the UML philosophy, a use case diagram could not be suitable for extracting the user interfaces. Use case diagrams may include some use cases referred to parts of the system not related to user interfaces such as classes, human tasks, components of other systems interacting with us, and so on. Or even
decomposing use cases by means of include and extend relationships, one could specify specific parts of the system, which are not related with the user interface. Therefore, in our opinion, a specialized version of the use case model could be required, or even some other UML diagrams could be used to complete the use case view.

Adapting UML to UI Design

Other works (Campos & Nunes, 2005; Conallen, 1999; Heumann, 2003; Lieberman, 2004; Nunes, 2003; Nunes et al., 2001; Pinheiro da Silva & Paton, 2000) propose the introduction of new UML elements in order to model user interfaces. In this case, the need of new stereotypes or classifiers in UML in order to distinguish GUI components from other system components is well known. In this case, there is a common point of view: classes should be stereotyped with <<screen>>, <<applet>>, and <<input_form>>, to be distinguished from data classes.

In addition, there are some works (Anderson, 2000; Elkoutbi & Keller, 2000) (Elkoutbi, Khriss, & Keller, 1999) interested in the description of the logic of GUI components, using activity and state diagrams for that. In this case, user interaction can be specified by means of states, where the system shows output data and requests input data. It should be specified how the user can navigate through the user interface, opening, closing windows, picking up in links, and so on. The specified logic should describe the set of states (input, output, windows) in which the user can be found.

Code Generation and UML

With respect to code generation, unfortunately, the capabilities of code generator CASE tools to transform design to an implementation are often restricted to produce class definitions consisting of attributes and operation signatures captured in class diagrams, but not methods to implement the procedural flow within the operations. Existing approaches in this last sense turn statecharts into executable code. Statecharts are used as object controllers for specifying when an object is willing to accept request. CASE tools supporting code generation from statecharts are Statemate (Ilogix, 2006a), Omate (Harel & Gery, 1997), Rhapsody (Ilogix, 2006b), and Fujaba (Schäfer, 2002).

MAIN THRUST OF THE CHAPTER

With regard to previous works on UML (use cases) and UI design, our contribution can be summarized as follows:

• Firstly, we consider use cases as a starting point of the user interface design. Here we define a new kind of diagram, user-interface diagram: a specialized version of the use case diagram as a high-level description of the user interface.

• In addition, following the MDD perspective, we integrate this system view with a set of specialized activity diagrams for user interaction design (user-interaction diagrams). One of the benefits of this integration is that each use case in the specialized use case diagram is described by means of a specialized activity diagram, and therefore interfaces can be analyzed and built from these two models.

• In addition, a class diagram is generated from the specialized activity diagrams. GUI components can also be built from both modelling techniques.

With regard to previous works on code generation, the UI models can get prototypes of the user interface of our application. Through mapping between UML and Java, we are able to generate low-level Java code directly from the user interaction diagram. This code generation is adapted to the special case of user interfaces, which is user event-based and handles input and output data by means of special kinds of UI components.
ORGANISATION OF THE CHAPTER

The next section describes our model-driven development technique for user interfaces. The chapter continues with the technique for the code generation by using user-interface models. This chapter finishes with some conclusions and future work.

MODEL-DRIVEN DEVELOPMENT FOR USER INTERFACES

Use Case Diagrams

Use case diagrams are used as starting point for user-interface design. Use cases are also a way of specifying required usages of a system, and they are typically used for capturing the requirements of a system (that is, what a system is supposed to do). The key concepts associated with the use-case model are actors and use cases. The users and systems that may interact with the system are represented by actors. Actors always model entities that are outside the system. Use cases represent the set of tasks that the actors carry out. In addition, the “use cases” can be decomposed by means of include relationships, and they can also be related by means of generalisation/specialisation relationships that compare more general and particular tasks.

In order to design a prototype of the user interface, the use case diagram should include the system actors and the set of (main) tasks for each one in which he or she takes part. From a point of view of user interface modelling, the use case diagram can be seen as a high-level description of the main windows of the system.

To illustrate the functionality of the MDD-based technique we will explain a simple Internet book shopping (IBS) model.

In the IBS example (Figure 1), three actors appear: the customer, the ordering manager, and the administrator. A customer directly makes the purchases on the Internet, querying certain issues of the product in a catalogue of books before making the purchase. The manager deals with customer’s orders (total or partially). And finally, the system’s administrator can manage the catalogue of books by adding and eliminating books in the catalogue or modifying those already existing. The administrator can also update or cancel certain component characteristics of an order or those orders fulfilling certain searching criteria.

This information is described with a use case diagram containing the identified actors and main tasks of the system. In our case study (see Figure 1), the actors are the customer, the manager, and the administrator, and the main tasks are purchase, manage orders, manage partial orders, manage catalogue, update orders, and update partial orders.

From this use case diagram, one can identify the future windows of the system that will comply with the needs of the presentation logic (graphical user interfaces).

User Interaction Diagrams

The second modelling technique in our framework is the activity diagram. However, we need to specialise the activity diagram for user interface design in the following sense.
Our activity diagrams include states and transitions. The states represent data output actions, that is, how the system responds to user interactions showing data (or requesting them). Then the user can introduce data and the corresponding event is handled and specified by means of transitions. Transitions can be conditioned, that is, the handled event is controlled by means of condition, which can be referred to data/business logic or a previous user interaction. In other words, it is possible more than one transition from a state, and to know which of them will run depends on data/business logic or the previous user choices. We call user-interaction diagrams to this kind of activity diagrams used for user interaction description.

Now, it is supposed that each use case in the use case diagram is described by means a user-interaction diagram. However, from a practical point of view, it is convenient to use more than one user-interaction diagram for describing a use case. This is so because the logic of a use case is usually too complex. For this reason, a user-interaction diagram can be deployed in several user-interaction diagrams, where a part of the main logic is separately described. For this reason, the main user-interaction diagram can be composed of three states (i.e., “query catalogue,” “confirm proceed,” and “shopping cart”). Two of them are terminal states (i.e., “confirm proceed” and “query catalogue”). A terminal state is described in a new diagram whose states correspond to graphical elements stereotyped (for instance <<JTextField>>, <<JList>>, and <<JLabel>> stereotypes. For instance, let us focus our attention in the purchase use case. Figure 2 shows the whole user-interaction diagram modelled for the purchasing process. Initially, the behavioural description starts with an original user-interaction description (Figure 2a). The behaviour shows how the customer begins the purchasing process of querying, adding, or removing articles of the shopping cart. After a usual purchasing process, the shopping system requests the customer’s card number and an address to carry out the shipment whenever the shopping cart is not empty. This diagram shows the graphical and behavioural content of the applet window where the purchases can be carried out.

The main user-interaction diagram (Figure 2a) is composed of three states (i.e., “query catalogue,” “confirm proceed,” and “shopping cart”). Two of them are terminal states (i.e., “confirm proceed” and “query catalogue”). A terminal state is described in a new diagram whose states correspond to graphical elements stereotyped (for instance <<JTextField>>, <<JList>>, and <<JLabel>>) and labelled by a text related to the graphical element. The name of a separate user-interaction diagram should be the same as that of the state. A non-terminal state is also described in a separate user-interaction diagram containing one or more non-terminal states.

Activity diagrams describe input and output user interactions. Given that we have decided to implement our user interface by means of Java swing package, we will consider the JFrame class as a container class that opens new frame windows (if needed). In addition, graphical components can be classified as input (a text field, a button, etc) and output components (a label, list, etc). Input/output components are associated with terminal states and transitions by using the appropriate stereotype. For instance, the stereotypes JTextField, JList, JLabel are associated with states and the stereotype JButton with transitions. Since the graphical behaviour concerns with states and transitions, next we will describe them separately.

States can be stereotyped or not. Stereotyped states represent terminal states, which can be labelled by <<JTextField>>, <<JList>>, and <<JLabel>> stereotypes. For instance, let us focus our attention in the purchase use case. Figure 2 shows the whole user-interaction diagram modelled for the purchasing process. Initially, the behavioural description starts with an original user-interaction description (Figure 2a). The behaviour shows how the customer begins the purchasing process of querying, adding, or removing articles of the shopping cart. After a usual purchasing process, the shopping system requests the customer’s card number and an address to carry out the shipment whenever the shopping cart is not empty. This diagram shows the graphical and behavioural content of the applet window where the purchases can be carried out.

Transitions can be labelled by means of stereotypes, conditions, or both. For instance, a button is connected with a transition by using a <<JButton>> stereotype, and the name of the label is the name of the button. For example, a show cart transition stereotyped as <<JBut-
Conditions can represent user choices or business/data logic. The first one is a condition of the user’s interaction with a graphical component (related to button or list states), and the second one is an internal checking condition (not related to the states, but to the internal process). For example, in our case study the selections from a list are modelled by conditions. Note in the query catalogue user-interaction diagram how the results list is modelled by a <<JList>> state and a [Selected article] condition.

Figure 2 shows some transitions (i.e., [close], [exit], or [proceed]) that correspond with conditions of the user choice type. The [Exit] output transition of the state query catalogue (Figure 2a) means that the user has pressed a button called exit, which has been defined in a separate query catalogue user-interaction diagram. Nevertheless, conditions like the [cart no empty] condition are business/data logic conditions, in which the human factor does not take part.

Furthermore, stereotyped transitions (buttons in our example) and conditions connect (non) terminal states with (non) terminal states. A condition would be an output of a non-terminal state if the user interacted with a button or a list component inside the respective non-terminal state. The usual way of “condition/event” transition can connect (no)terminal states with (non) terminal states. A condition/event transition between states means which condition should be present to trigger the event. In our case study, an event can only be a button. For instance, to remove an article from the shopping cart, it must previously be selected from the cart list (Figure 2c).

User Interface Diagrams

Therefore, in our technique we have obtained a use case diagram together with a set of user-interaction diagrams, from which some correspond to use cases and others to states of use cases. How-
ever, it could be useful to have a new version of the use case diagram, to know what are the main user-interaction diagrams, that is, which are the user-interaction diagrams corresponding to use cases, and which are the secondary user-interaction diagrams, that is, which are states of use cases. For this reason, we will build a new version of the use case diagram called user-interface diagram as follows. The user-interface diagram contains the same actors and use cases of the use case diagram. In addition, we will add states (as use cases) of each use case corresponding to user interaction diagrams. In order to connect use cases, we will use the “include” or “generalisation” use case relations.

User Interface Relationships

Include and generalisation relationships have the following meaning in this user-interface diagram, similarly to the same relations in the use case diagram. If a use case contains in its user interaction diagram a state described by means of another user interaction diagram, the state is also considered as a use case, and the included state (use case) is related in the user interface diagram by the inclusion relationship with the main use case. However, this is achieved whenever the logic of the included use case is not modified by the main use case. The logic of an included use case is modified in the main use case whenever the user interaction diagram of the main use case includes transitions that are referred to internal states or conditions of the included use case. This typically happens when the main use case defines its own logic using the included use case as a piece of behaviour, but access to the included use case. Otherwise, that is, whenever the logic of the included use case is modified, we relate the main use case and the state (use case) by means of the generalisation relationship.

In addition, we have to identify a special case of generalisation relationship, in which we have a main use case including a state and the state itself is a specialisation version of another use case. This happens when the user interaction diagram of the state specialises the user interaction diagram of another use case. The specialisation consists of the replacement of states and transitions of the user interaction diagram by means of more particular states and transitions with similar semantics. Here we are interested in the identification of similarity between interactions following the same pattern; however, the difference in our case consists of several kinds of GUI components and input/output data.

In other words, the specialisation of use cases allows us to build new use cases with a more complex logic containing the specialized use case, and adding transitions and states or modifying the existent ones. On the contrary, the inclusion allows us to build new use cases with a more complex logic without adding or modifying the states and transitions of the included use case.

Given that the user interface diagram is a kind of high-level description of the user interface, the developer still could decide not to include in it all the states representing user interaction diagrams. In other words, the user interface dia-
gram represents the set of windows of the system by means of use cases. However, there could be more user interaction diagrams than windows. Usually, some user interaction diagrams can be built for deploying some states but they will not correspond to system windows.

Once user-interaction diagrams have been described, the designer proceeds to build the user interface diagram. It contains new use cases that are some of the non-terminal states of the user interaction diagrams. In addition, the developer has to identify use case relationships in the new user interface diagram as follows.

Include Relationships

Let us consider the purchasing process described in previous user-interaction diagrams. The purchase use case is a frame that includes (uses) three other frames (use cases): query catalogue, shopping cart, and confirm proceed. These use cases are described by means of separate user interaction diagrams. In addition, the logic of the use cases is not modified in the purchase user interaction diagram. It integrates the logic query catalogue diagram by checking which buttons (i.e., exit) the user pressed when he or she exits from shopping cart.

The developer can also identify an inclusion relationship between manage catalogue and withdraw article, Modify article and add article use cases (Figure 4). In these cases, four windows can be optionally opened (depending on a menu) from the manage catalogue window. In addition, the administrator identification window is mandatory opened from the Manage catalogue window in order to achieve the system's administrator tasks.

Generalisation Relationships

In order to illustrate the generalisation/specialisation relationship, we will pay attention to three use cases: purchase, query catalogue, and query catalogue by administrator. In previous sections, we have identified two cases of generalisation/specialisation.

The first case is the query catalogue and purchase case. In this case, the purchase user-interaction diagram contains a state (use case) that specialises query catalogue in the following sense. The query catalogue user interaction diagram describes how to query the catalogue of the IBS by introducing the searching criteria and showing the results. However, the purchase user interaction diagram can interrupt
the querying process by adding the searched items to the shopping cart. It is specified by adding the Add to cart button as transition from (and to) query catalogue. Therefore we can identify a specialisation relationship between purchase and query catalogue. It is also supposed that there will be a window for query catalogue from which purchase inherits.

The second case is the relationship between query catalogue and query catalogue by administrator. In this case, the administrator is supposed to have higher privileges for querying the catalogue and therefore the user interaction diagram of the query catalogue by administrator (see Figures 5 and 6) specialises in the query catalogue user interaction diagram in the following sense. The states of the query catalogue by administrator corresponding with the searching criteria and results are modified with respect to the query catalogue. It is supposed that the searching and result fields are different but the logic is similar. In other words, the query catalogue can be replaced by query catalogue by administrator given that the states can be replaced. In this case, we can identify a generalisation relationship between them. Analogously, Withdraw article and Modify article combine both kinds of specialisation, once they’ve specialized the query catalogue by administrator in the same sense that purchase specializes query catalogue, and they specialize indirectly query catalogue (see Figure 6).

The complete user interface diagram of our case study can be seen in Figure 7.

**GUI Class Diagram**

The next step of our model-driven technique consists of the building of a class diagram for GUI components. The user-interface diagrams obtained in the previous state give us the main windows. Each use case connected to an actor can be converted into a window, and if an actor
is connected to more than one use case, it can be considered a window by the actor that invokes (or embeds) each window of each use case.

Therefore, the actor window can be a menu window. In addition, in the user-interaction diagrams obtained from use cases, we have also described input and output components for data output and request and user events. It gives us the GUI components for each window. If a user-interaction diagram has a state described by means of another user-interaction diagram, we can suppose that the window of the use case could also contain a separate window for this separate task. However, now, we have to take into account the user-interface diagram and the use case relationships. In the case of inclusion, the logic is also separate, and it is possible to consider a new window. However, in the case of generalisation/specialisation, the window corresponds with a specialisation, and therefore it is better to consider a new window by using the inheritance relation.
into account the user-interface diagram and the use case relationships. In the case of inclusion, the logic is also separate, and it is possible to consider a new window. However, in the case of generalisation/specialisation, the window corresponds with a specialisation, and therefore it is better to consider a new window by using the inheritance relation.

The windows obtained from use cases and the GUI components obtained from each user-interaction diagrams allow us to build a class diagram, which we call GUI-class diagram, where there are windows classes connected by associations or by inheritance. In addition, there are GUI component classes connected by associations to window classes.

Once the user-interface diagram has been built and a set of user-interaction diagrams has been obtained, now we can generate a GUI-class diagram. The GUI-class diagram is built from Java swing classes. In the method, each use case corresponds with a frame class. Use cases are translated into classes with the same name as these use cases. The translated classes specialise a Java Frame class. The components of the frame (use case) are described in user-interaction diagrams. A terminal state is translated into that Java swing class represented by the stereotype of the state. The Java swing class is connected from the container class. For example, those terminal states stereotyped as <<JTextField>> are translated into a JTextField class in the GUI-class diagram. Something similar happens with the rest of stereotyped states and transitions. Figures 8 to 12 show the main classes of GUI-class diagram of the customer’s side.

As it can be seen in these figures, the stereotyped states and transitions in the user-interaction diagrams are translated into Java classes in the GUI-class diagram. The stereotype name of a transition or state is translated into the appropriate Java swing class. For example, the <<JButton>> stereotype of the Proceed transition that appears in the purchase user-interaction diagram (see Figure 2a) is translated into a JButton class.
The windows obtained from use cases and the GUI components obtained from each user-interaction diagrams allow us to build a class diagram, which we call GUI-class diagram, where there are windows classes connected by associations or by inheritance. In addition, there are GUI component classes connected by associations to window classes.

Once the user-interface diagram has been built and a set of user-interaction diagrams has been obtained, now we can generate a GUI-class diagram.

The GUI-class diagram is built from Java swing classes. In the method, each use case corresponds with a frame class. Use cases are translated into classes with the same name as these use cases. The translated classes specialise a Java Frame class. The components of the frame (use case) are described in user-interaction diagrams. A terminal state is translated into that Java swing class represented by the stereotype of the state. The Java swing class is connected from the container class. For example, those terminal states stereotyped as <<JTextField>> are translated
into a JTextField class in the GUI-class diagram. Something similar happens with the rest of stereotyped states and transitions. Figures 8 to 12 show the main classes of GUI-class diagram of the customer’s side.

As it can be seen in these figures, the stereotyped states and transitions in the user-interaction diagrams are translated into Java classes in the GUI-class diagram. The stereotype name of a transition or state is translated into the appropriate Java swing class. For example, the <<JButton>> stereotype of the Proceed transition that appears in the purchase user-interaction diagram (see Figure 2a) is translated into a JButton class.

GUI Prototypes

Finally, rapid GUI prototypes could be obtained from the GUI-class diagram. Figure 13 shows a first visual result of the purchase window. Note how the purchase window is very similar to the query catalogue window, except that the second one includes three buttons more than the first window. This similarity between windows was revealed in the user interface diagram as a generalisation relationship between use cases: between the query catalogue and purchase use cases. In the IBS design, the customer will always work on a purchase window opened from the Customer window, and never on a query catalogue window, though the former inherits the behaviour of the latter (i.e., by the relation of generalisation). Let us remark that purchase inherits from query catalogue, and that the five windows (GUI) are the five use case of the client side (see Figure 3).
The shopping cart window (Figure 13c) appears when the show cart button is pressed on the purchase window (Figure 13b). Note in the user interface diagram, shown in Figure 7, how the button is associated with the window by means of an inclusion relation between use cases. On the other hand, the two information windows (Figure 13d) are also associated with two buttons: the remove article button in the shopping cart window and the proceed button in the purchase window. Note again how these windows are also described as inclusion relations between use cases.

Moreover, observe the user-interaction diagrams shown in Figure 2 to better track the behaviour of the example. To develop the example, we have used the Rational Rose for Java tool. For space reasons, we have only included a part of the GUI project developed for the case study. A complete version of the project is available at http://indalog.ual.es/mdd/purchase.

CODE GENERATION FOR USER INTERFACES

Finally, our model driven technique allows us to obtain code generation from the developed models. The user-interaction diagrams describe each user interaction of each use case by means of states representing data output and transitions representing data input and events.

Code generation consists of coding an event-based user interface. We have decided to implement our user interface by means of the Java swing classes, and therefore we will use Applet and Frame classes for window components, and button, label, list, etc., for other UI components. According to the Java swing package, events are handled by means of listeners and event handlers in the corresponding window classes.

In order to explain the transition of the user interaction diagrams to code, we should first track a real interaction of the users with windows from the example: the purchase behaviour from the client side.

Here we can see the complete sequence of user/windows interaction. The first window (purchase, slide 1) shows ten graphical components: two labels (searching criteria and results), one text field (textfield), one field with a list of results (results), and six buttons. As we can see, the search, close and exit buttons are enabled whereas buttons add to cart, show cart, and proceed are not active. The user can only choose one searching criteria in the corresponding field and then click search. This interaction with the window after clicking the button makes the user access the logic data to obtain the results and then he or she gets to slide 2 window.

In that window (slide 2), we can observe that the searching button has been deactivated and the user can choose either one element from the list or exit the application or clear the searching criteria, in which case the search button is again activated. Following the usual interaction sequence, when the user is in the slide 2 window, he or she can choose one item from the list and then he or she will go to slide 3.

Moreover in slide 3, one button to add the cart is activated. The user adds to the cart the selected element by pressing the button and then he or she gets to the slide 4 window. In this state, there are some buttons that can show, on the one hand, the content of the cart (show cart) and on the other hand, they can make the purchase (proceed). The user can choose different elements from the list and add them to the cart. He or she can also enter new elements to the cart from a new searching, clearing the list (clear), and entering a new “searching criteria” (slide 5). In this state, we can observe that the searching button is activated and the adding to cart button is deactivated. However, show cart and proceed buttons are not activated because the cart is not empty and the user at any time can see the cart or make the purchase. If the user wishes to see the content of the cart, he or she would get to the slide 6 window (shopping cart).
User Interaction and Interface Design with UML

Figure 14. All complete sequence of the purchase interaction

![Sequence of purchase interaction](Figure14.jpg)

This new window (slide 6) has two labels, one list and two buttons; the first one is deactivated to avoid the removal of an element from a list until the user hasn’t selected one of them and the second one allows us to close the window.

If the user selects one element from the list (slide 7), the button to eliminate it from the cart is activated. If so, he or she would get to the third level of the window for requesting the user’s confirmation to eliminate the element from the cart (slide 8).

If the user confirms the elimination (in window slide 8) he or she would get to the last level of the window shown in slide 9; otherwise, he or she would get to the slide 7 again.

General Structure of Frame Classes

The interaction behaviour previously described is defined by the designer through user-interaction and user-interfaces diagrams. The relationship of dependences between windows is described.
Figure 15. A summary for the user-interface and user-interaction diagram of the purchase function

by means of user-interfaces diagrams (see again Figure 3).

The behaviour of the window when the user activates and deactivates buttons and changes the state of the graphical components is described by means of user-interaction diagrams (see again Figure 2). In order to obtain the behaviour shown in Figure 14, modelling it with user-interface and user-interaction diagrams (Figures 2 and 3), we create code prototypes in Java from these diagrams, making them correspond with specific lines of Java code. The most general structure of our method is the Frame classes for the windows of the system.

To explain how to make this relationship between diagrams and code, we are focusing on the purchase user-interaction diagram. Figure 15 shows three perspectives of the purchase:

1. The user-interface perspective, where the designer studies the dependences between windows through use cases;
2. The user-interaction perspective, where the designer models the behaviour of the graphic components of the windows and their interaction with the user; and

3. The GUI perspective, where the user works with a rapid prototype of the windows through a rapid generation of the code from both aforesaid perspectives (as previously seen).

In a symbolic way, we have represented in the picture the GUI perspective (of the windows) with its equivalent in the user-interface perspective, including the type of relationship between windows (in order to clarify the dependences between them).

In this example, these five use cases lead to five Java files of frame type. Their names are obtained from the use case name (which will be the same as the one for the user-interaction diagram, one per use case). If compound nouns exist then a simple name is created with the initials of each word in capital letters (Table 1).

Classes extend JFrame class and present a general structure of implementation based on frames (see Table 2). The basic structure of frame prototype created in the diagrams-to-code transition, is composed of four sections: one heading, one class constructor, one initiator of the frame (jbInit) and one basic prototype of the implementation of the detected methods (related with the logic of presentation). In the heading (labelled with /** @GUIcomponents */) the attributes are considered to be graphical components that have been detected in the corresponding user-interaction diagram. The class constructor is the same for all code files created (except for the name of the constructor). In the initiator of the frame (jbInit) it is included the code lines that model the graphical content of the window. The designer should establish the position and size of the graphical components of the frame later on, as the code generated is just a prototype of the window. In the fourth section of the Frame, it is created the basic methods related to the interaction of the user with the graphic elements that he or she can interact with: buttons, lists, and text fields.

**GUI Components**

In order to analyze how translating the diagrams components to code inside the frame, we’ll again use the purchase example in Figure 15.

The behaviour of the purchase window begins with the initial state query catalogue (represented
by the designer as an inheritance of behaviour in the user-interface diagram. As it can be observed, the purchase window inherits all the graphical components (and behaviour) of the “query catalogue” window and it also adds the add to cart, show cart and proceed buttons, which are not activated. From the initial state query catalogue, the four out coming transitions correspond to interactions with the buttons of the window and deal to new states. What the diagram shows from its initial state is the following: If there is an element selected from the list, then it is activated the button to add one element to cart.

```java
<<JButton>>
[ selected article ] / Add to cart
```

If the cart is not empty ([cart not empty]) it is activated the button to show the content of the cart and make the purchase of the products in the cart.

```java
<<JButton>>
[ cart not empty ] / Shopping cart
```

```java
<<JButton>>
[ cart not empty ] / Proceed
```

The transition [exit] in the initial query catalogue state, corresponds with the operation of clicking on the exit button inside the behaviour diagram of the window (frame): query catalogue (window inherited by purchase). The same happens with the other two transitions that reach this initial state from the state Confirm proceed (window to confirm the purchase).

For this example, there are four modelled graphical components: labels, text fields, lists, and buttons. In Table 3, it is shown the correspondence between graphic components of the user-interaction diagram (stereotypes) and the code generation.

In Table 3, column “sterotype” represents the stereotypes used in the states/transitions of the user-interactions diagram.

The columns “attribute” and “class” represent the name of the attribute and the base class instantiated in the code generated. The criterion followed to establish the name generated in the code is: name_Type. That is, as a name it is used the same name indicated in the interaction diagram in capital letters, followed by a hyphen (i.e., “_”) and then finished by base type (label, button, list, etc.). If the original name has blanks

<table>
<thead>
<tr>
<th>UI diagrams</th>
<th>Code generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereotype</td>
<td>Attribute</td>
</tr>
<tr>
<td>&lt;&lt;JLabel&gt;&gt;</td>
<td>name</td>
</tr>
<tr>
<td>&lt;&lt; JTextField&gt;&gt;</td>
<td>name</td>
</tr>
<tr>
<td>&lt;&lt;JList&gt;&gt;</td>
<td>name</td>
</tr>
<tr>
<td>&lt;&lt; JButton &gt;&gt;</td>
<td>name</td>
</tr>
<tr>
<td>[condition]</td>
<td>None</td>
</tr>
</tbody>
</table>
in between, they are substituted by hyphens. For example, the label:

```
<<JLabel>>
```

Shopping cart

It is translated to the name of the attribute shopping cart button and then it is generated the following Java code:

```java
public JLabel shopping_cart_Button = new JLabel();
```

The “in/out” column represents an exit from/to the user (or window). So, for example, one stereotype label will represent a text shown from the system towards the user (window) whereas the button (according to its graphic behaviour) represents an income action (interaction) from the user towards the system. The “textfields” and lists stereotypes represent both options: the system shows and modifies the content of these graphic components and at the same time, the user can either write (textfield) or choose (list) in these graphic components. The “out” components have no correspondence with the code, whereas the “in” components have correspondence with one listener class.

The “listener” column indicates whether a graphical component is moved by means of a listener class. This type of class allows the class creating it for listen the behaviour of the graphic component with the user. It is usual to use listener in buttons and lists (we’ll study these later).

The “S/T” column is referred to whether the graphic component corresponds with a state or a transition in the user-interaction diagram.

Finally, “markup” column is referred to the mark prototype created before the translated code. This mark helps to identify the content of the text when the programmer has to rewrite or insert the new code to add and complete the functionality of the windows.

### Mapping States into Code

As previously said, the states of a user-interaction diagram can be referred to terminal states (label, textfield, list) and non-terminal states

#### Non-Terminal States

Let us begin with the user-interaction diagram purchase (see Figure 15). The diagram is composed of three non-terminal states. We should remember that a non-terminal state is a state
For a label case, three basic code lines are generated: one in the heading to define the type (JLabel) and two in the initiation (jbInit) to establish the component. In Figure 16, it is shown a bit of code generated to the label state "confirm remove article." The code line 3 creates one variable of JLabel type. Note how the name of the variable is obtained from the name given to the <<JLabel>> state in the user-interaction diagram. The line 7 establishes the text that will appear in the window. Line 9 adds the label to the container. Lines 2, 6, and 9 are marks established previously that are inserted in the code to facilitate the identification of each part of the code.

Table A.2 of Appendix A shows the Java code pattern that is translated from the <<JLabel>> graphical components of user-interaction diagram.

For the "text fields," five basic code lines are generated. Two of them are generated in the heading to define two graphical components in the heading: one JTextField and one JLabel to name it in the window. The other three code lines are generated in the initiation section of Figure 18. The results list state in query catalogue

```java
1: public class QueryCatalogue extends JFrame {
2:   /** @GUIcomponents */
3:   public JLabel searching_criteria_Label = new JLabel();
4:   public JTextField searching_criteria_TextField = new JTextField();
5:   
6:   private void jbInit() throws Exception {
7:     /** @Label */
8:     name_Label.setText("Searching criteria");
9:     
10:    /** @Panel */
11:    this.getContentPane().add(searching_criteria_Label, null);
12:    this.getContentPane().add(searching_criteria_TextField, null);
13:  } // end jbInit()
14: }
```

```java
1: import javax.swing.event.*;
2: 
3: public class QueryCatalogue extends JFrame {
4:   /** @GUIcomponents */
5:   public JLabel results_Label = new JLabel();
6:   public JScrollPane results_ScrollPane = new JScrollPane();
7:   public JList results_List = new JList();
8:   
9:   private void jbInit() throws Exception {
10:     /** @Label */
11:     results_Label.setText("Results");
12:     
13:     /** @List */
14:     results_ScrollPane.getViewport().add(results_List, null);
15:     results_List.addListSelectionListener(
16:         new javax.swing.event.ListSelectionListener () {
17:             public void valueChanged(ListSelectionEvent e) {
18:                 results_List_ShoppingCart_GUI(); }
19:         });
20:     
21:     /** @Panel */
22:     this.getContentPane().add(results_Label, null);
23:     this.getContentPane().add(results_ScrollPanel, null);
24:  } // end jbInit()
25: 
26: // Methods ...
27: void results_List_ShoppingCart_GUI() { ... }
28: }
```
whose behaviour is defined separately from other diagram of interaction. For example, the query catalogue, confirm proceed, and shopping cart states will be defined in independent user-interaction diagrams having the same name as the corresponding state. In the user-interface diagram, the query catalogue is established by the designer as an inheritance for purchase, whereas the other two states (which are “use cases” and therefore windows) are established as conclusions in the user-interface diagram (i.e., <<include>>). The translation is as follows:

```java
public class purchase extends querycatalogue {
    /** @GUI */
    ShoppingCart shoppingcart_GUI = new ShoppingCart();

    /** @GUI */
    ConfirmProceed confirmproceed_GUI = new ConfirmProceed();
    ...
```

Note that purchase class inherits querycatalogue in the code. The two “include” relationships are included in the code in the heading section (one of the four that frame has). On the code line a @ GUI mark is inserted to trace an inclusion of the window in the process of translation. Moreover, the criterion for the name of the GUI variable is considering the original name of the included window followed by “_GUI.”

Terminal States

Terminal states (label, list, textfield) are translated into code lines in the heading, initiation of Frame (jbInit) and implementation of methods sections (only in list). Appendix A contains the Java code patterns translated from the graphical components of the user-interaction diagram. Let’s see each of them separately.

For a label case, three basic code lines are generated: one in the heading to define the type (JLabel) and two in the initiation (jbInit) to establish the component. In Figure 16, it is shown
a bit of code generated to the label state “confirm remove article.” The code line 3 creates one variable of JLabel type. Note how the name of the variable is obtained from the name given to the <<JLabel>> state in the user-interaction diagram. The line 7 establishes the text that will appear in the window. Line 9 adds the label to the container. Lines 2, 6, and 9 are marks established previously that are inserted in the code to facilitate the identification of each part of the code.

Table A.2 of Appendix A shows the Java code pattern that is translated from the <<JLabel>> graphical components of user-interaction diagram.

For the “text fields,” five basic code lines are generated. Two of them are generated in the heading to define two graphical components in the heading: one JTextField and one JLabel to name it in the window. The other three code lines are generated in the initiation section of frame (jbInit). For example, Figure 17 shows how translating the terminal state searching criteria (in the user-interaction diagram of frame query catalogue) into a code in two components (lines 2 and 3). It is created a “text field” area (JtextField) and a text (label) which accompanies the text field. For the naming of the variables we use the same criterion established before: the same text that appears in the state followed by “label” or “textfield”. In line 9, it is established the text in the window. Lines 11 and 12 add both graphical components to the container.

Table A.3 of Appendix A shows the Java code pattern that is translated from the <<JtextField>> graphical components of the user-interaction diagram.

Finally, there are the terminal states of list type. Graphically these states are somehow special since—like a textarea—they can be translated into different ways: a label component, other list or even a scrollpane. The latter can be used as a container to show the elements from the corresponding list. Several lines are generated in the translation, in three sections (heading, initiation, and methods). For example, Figure 18 shows a translation of the results state.

Lines 5, 6, and 7 are generated in the heading (after /** @GUIcomponents */) defining the name of the variables of the three graphical components. Again the criterion used for the naming of the variables is the same as that one used in previous examples. In the initiation section (jbInit) the following code is created. Firstly, the label text is established in the panel (line 11). Then, an event of listener type is used to catch the interactions of the user with the list when he or she presses the elements of the list. Therefore, in line 14, there is the container JScrollPane type where the list is and from lines 15 to 18 there is the listener event on the list. To do so, the library javax.swing.event.* is imported (line 1). Lines 21 and 22 add graphical elements to the panel. Finally, the method implementation section (line 26) includes the prototype of the method used each time the user presses on an element from the list. The implementation of the method is a later task of the programmer.

In table A.3 of Appendix A it is shown the Java code pattern translated from JList graphical components of user-interaction diagram.

Mapping Transitions into Code

Transitions correspond with buttons (<<JButton>>) or changes in the internal state of the data logic. They are represented in the diagram as a condition markup (p.e., [cart not empty] o [Close]). Table A.1 (Appexd A) shows the Java code pattern for a graphical component <<JButton>> translated from the user-interaction diagram.

Transitions represent graphical components that generate code at several sections of the program. The translation into code is the more important issue in the code generation process, since it reveals the usefully of the user-interaction diagrams for the traceability of the interaction of the user with the graphical components of the
window and also how that (window) responds to the user. To explain the translation of states to code, we will use again the purchase windows sample. In Figure 19, we show (at the right side) the two related user-interaction diagrams of the purchase window (showed at left side).

Remember that the system’s designer has described the purchase window from the query catalogue window, which represents a state in the user-interaction diagram of the purchase window. This state inherits the inner behaviour of the query catalogue (whose behavioural interaction is shown at the right-lower side in Figure 19). It arrives at two possible states through four transitions from the state query catalogue: (a) see the state of the cart (shopping cart state); or (b) carry out the purchase of the products stored in the cart (confirm proceed state). Both states are reached from the purchase window through the buttons show cart and proceed. In the state, a conditioned transition has been included for reaching these two windows from purchase: that the cart is not empty. In other words, a conditioned button-transition corresponds with an operation of activation/deactivation of buttons.

Since these three buttons are conditioned by a selection action in a list, a listener operation is generated on the list in the purchase class. Remember that both the list and the most of the graphical components of the purchase window are inherited from query catalogue window (i.e., public class purchase extends querycatalogue), and therefore the operation listener is done on the superclass, which contains the list.

In this way, whenever the user presses on the list, an event is captured, which calls to the function selected_articles_list_querycatalogue_GUI(). The name of the method is established by the text included in the condition of the transition in the user-interaction diagram. For instance, a button-transition Add to cart containing a condition [selected articles] means that the class (frame) purchase enables the button Add to cart in the window if and only if an element of the list has been selected; in other case, the button will remain disabled (see Figure 19). In case of exist a selected element in the list, the button is activated and the transition leads again to the same state, to include the selected element (and others) to the list. This generates the following lines of code for the method on the listened element.

```java
/** @List */
public class purchase extends querycatalogue {
    /** @GUIcomponents */
    public JButton add_to_cart_Button = new JButton();

    private void jbInit() throws Exception {
        /** @Button */
        name_Button.setText("Add to cart");
        name_Button.addActionListener( 
            new java.awt.event.ActionListener() { 
                public void actionPerformed(ActionEvent e) { 
                    add_to_cart_Button_Purchase_GUI(); 
                } });

        /** @Panel */
        this.getContentPane().add(add_to_cart_Button, null); 
    } // end jbInit()

    // Methods ...
    /** Button */
    void add_to_cart_Button_Purchase_GUI() {
        ... 
    }
}
```

The activation of the button (add_to_cart.Button.setEnabled) implies the button has been before declared in the panel, created and added.
Following the trace of the example, the other two button-transitions that leave the state *query catalogue* remain disabled until the cart is not empty (button *Add to cart*):

```java
/** @Button */
public void add_to_cart_Button_purchase_GUI() {
  show_cart_Button.setEnabled(true);
  proceed_Button.setEnabled(true);
}
```

Once the buttons is activated, two other states can be reached from the state *query catalogue*: (a) the state *shopping cart*, whether the button *show cart* is pressed; or (b) the state *confirm proceed*, whether the button *proceed* is pressed. Since both states are not stereotyped, they are translated into two independent windows (frames) with independent behaviours in *user-interaction diagrams*, also independent (i.e., with an initial state and another end state).

Nevertheless, this encapsulated behaviour (i.e., a *user-interaction diagram*) can be interrupted by means of conditioned simple transitions (or simply “simple transitions”). Simple transitions are those transitions that are labelled only by a condition (p.e., [screen], [accept], [close], and [exit]). These transitions refer to an action of pressing a button that exists inside the non-terminal state (a window). For instance, the button-transitions [accept] and [screen] refer that, being in the state (frame) *confirm proceed* the user presses one of these buttons, and one returns to the state (frame) *query catalogue*. The same for the transition [close].

Nevertheless, there is a special case of simple transition: that one that does not come from another state, but it departs from the target state. For example, the simple transition [exit] means that there has been pressed the button exit inside the frame *query catalogue*, whose behaviour is defined in another independent *user-interaction diagram* (in which the button is defined). In any case, both the simple-transitions of the type [accept] or of the type [exit] the interpretation in the code is reflected by means of a *listener* operation on those buttons of the window (class) containing the buttons, that interrupts the normal interaction sequence of the above mentioned class.

```java
public class purchase extends querycatalogue {
...
  private void jbInit() throws Exception {
...
    // Listener on children
    /** @Button */
    shoppingcart_GUI.close_Button.addActionListener(
        new java.awt.event.ActionListener() {
          public void actionPerformed(ActionEvent e) {
            close_Button_ShoppingCart_GUI();
          }
        });
  }
  /** @Button */
  confirmproceed_GUI.accept_Button.addActionListener( ...);
  /** @Button */
  confirmproceed_GUI.cancel_Button.addActionListener( ...);
}
```

Appendix B partially shows the java code of one of the five classes of the project example: the class *ShoppingCart*. This information is available in the following Web page: http://indalog.ual.es/mdd/purchase.

**CONCLUSION AND FUTURE WORK**

In this chapter, we have studied a model-driven development technique for GUI design. We have defined a new kind of UML diagram (i.e., the *user-interface diagram*), which specialises in the use case diagram for user interface design. In addition, we have shown how to describe use cases by means of specialized activity diagrams (*user-interaction diagrams*) in order to specify the user interface of each use case. Finally, we have shown how to generate class diagrams for GUI and how to build rapid prototypes for the user interface. As future work, we plan to study the following extensions of our work:
1. Integrate the user interface view with the data and business logic;
2. Incorporate our method in a CASE tool in order to automate it;
3. Integrate our technique in the whole development process.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish project of the Ministry of Science and Technology TIN2005-09207-C03-02 and TIN2006-06698 under FEDER funds.

REFERENCES


### APPENDIX A. GUI CODE PATTERNS

#### A.1. Mapping transitions (buttons) into code

<table>
<thead>
<tr>
<th>User-Interaction Stereotype</th>
<th>Code Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;&lt;JButton&gt;&gt; name</code></td>
<td><code>import java.awt.event.*;</code>&lt;br&gt;<code>public class className extends JFrame {</code>&lt;br&gt;<code>/** @GUIcomponents */</code>&lt;br&gt;<code>public JButton name_Button = new JButton();</code>&lt;br&gt;<code>private void jbInit() throws Exception {</code>&lt;br&gt;<code>/** @Button */</code>&lt;br&gt;<code>name_Button.setText(&quot;name&quot;);</code>&lt;br&gt;<code>name_Button.addActionListener(new java.awt.event.ActionListener() {</code>&lt;br&gt;<code>public void actionPerformed(ActionEvent e) {</code>&lt;br&gt;<code>name_Button._nameClass_GUI();</code>&lt;br&gt;<code>}});</code>&lt;br&gt;<code>/** @Panel */</code>&lt;br&gt;<code>this.getContentPane().add(name_Button, null);</code>&lt;br<code>} // end jbInit()</code>&lt;br&gt;<code>/** * Method THIS */</code>&lt;br&gt;<code>/** Button */</code>&lt;br&gt;<code>void name_Button._nameClass_GUI() { ... }</code>&lt;br&gt;<code>}</code></td>
</tr>
</tbody>
</table>

#### A.2. Mapping labels to code

<table>
<thead>
<tr>
<th>User-Interaction Stereotype</th>
<th>Code Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;&lt;JLabel&gt;&gt; name</code></td>
<td><code>public class className extends JFrame {</code>&lt;br&gt;<code>/** @GUIcomponents */</code>&lt;br&gt;<code>public JLabel name_Label = new JLabel();</code>&lt;br&gt;<code>private void jbInit() throws Exception {</code>&lt;br&gt;<code>/** @Label */</code>&lt;br&gt;<code>name_Label.setText(&quot;name&quot;);</code>&lt;br&gt;<code>/** @Panel */</code>&lt;br&gt;<code>this.getContentPane().add(name_Label, null);</code>&lt;br<code>} // end jbInit()</code>&lt;br&gt;<code>/** * Method THIS */</code>&lt;br&gt;<code>/** Button */</code>&lt;br&gt;<code>void name_Label._nameClass_GUI() { ... }</code>&lt;br&gt;<code>}</code></td>
</tr>
</tbody>
</table>

#### A.3. Mapping textfields to code

<table>
<thead>
<tr>
<th>User-Interaction Stereotype</th>
<th>Code Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;&lt;JTextField&gt;&gt; name</code></td>
<td><code>public class className extends JFrame {</code>&lt;br&gt;<code>/** @GUIcomponents */</code>&lt;br&gt;<code>public JLabel name_Label = new JLabel();</code>&lt;br&gt;<code>public JTextField name_TextField = new JTextField();</code>&lt;br&gt;<code>private void jbInit() throws Exception {</code>&lt;br&gt;<code>/** @Label */</code>&lt;br&gt;<code>name_Label.setText(&quot;name&quot;);</code>&lt;br&gt;<code>/** @Panel */</code>&lt;br&gt;<code>this.getContentPane().add(name_Label, null);</code>&lt;br&gt;<code>this.getContentPane().add(name_TextField, null);</code>&lt;br<code>} // end jbInit()</code>&lt;br&gt;<code>}</code></td>
</tr>
</tbody>
</table>
### A.4. Mapping lists to code

<table>
<thead>
<tr>
<th>User-Interaction Stereotype</th>
<th>Code Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;&lt;JList&gt;&gt; name</td>
<td>import javax.swing.event.*;</td>
</tr>
<tr>
<td></td>
<td>public class className extends JFrame {</td>
</tr>
<tr>
<td></td>
<td>/* @GUIcomponents */</td>
</tr>
<tr>
<td></td>
<td>public JLabel name_Label = new JLabel();</td>
</tr>
<tr>
<td></td>
<td>public JScrollPane name_ScrollPane = new JScrollPane();</td>
</tr>
<tr>
<td></td>
<td>public JList name_List = new JList();</td>
</tr>
<tr>
<td></td>
<td>private void jbInit() throws Exception {</td>
</tr>
<tr>
<td></td>
<td>/* @Label */</td>
</tr>
<tr>
<td></td>
<td>name_Label.setText(&quot;name&quot;);</td>
</tr>
<tr>
<td></td>
<td>/* @List */</td>
</tr>
<tr>
<td></td>
<td>name_ScrollPane.getViewport().add(name_List, null);</td>
</tr>
<tr>
<td></td>
<td>name_List.addListSelectionListener(</td>
</tr>
<tr>
<td></td>
<td>new javax.swing.event.ListSelectionListener() {</td>
</tr>
<tr>
<td></td>
<td>public void valueChanged(ListSelectionEvent e) {</td>
</tr>
<tr>
<td></td>
<td>selected_articles_List_ShoppingCart_GUI(); } });</td>
</tr>
<tr>
<td></td>
<td>/* @Panel */</td>
</tr>
<tr>
<td></td>
<td>this.getContentPane().add(name_Label, null);</td>
</tr>
<tr>
<td></td>
<td>this.getContentPane().add(name_ScrollPane, null);</td>
</tr>
<tr>
<td></td>
<td>} // end jbInit()</td>
</tr>
<tr>
<td></td>
<td>/**</td>
</tr>
<tr>
<td></td>
<td>* Method THIS</td>
</tr>
<tr>
<td></td>
<td>*/</td>
</tr>
<tr>
<td></td>
<td>/** @List */</td>
</tr>
<tr>
<td></td>
<td>void selected_articles_List_ShoppingCart_GUI() { ... }</td>
</tr>
<tr>
<td></td>
<td>}</td>
</tr>
</tbody>
</table>
**User Interaction and Interface Design with UML**

### APPENDIX B. THE SHOPPING CART FRAME

```java
public class ShoppingCart extends JFrame {

    /** @GUIcomponents */
    public JLabel shopping_cart_Label = new JLabel();
    public JLabel selected_articles_Label = new JLabel();
    public JButton remove_article_Button = new JButton();
    public JButton close_Button = new JButton();
    public JScrollPane selected_articles_ScrollPane = new JScrollPane();
    public JList selected_articles_List = new JList();

    /** @GUI */
    ConfirmRemoveArticle confirmremovearticle_GUI = new ConfirmRemoveArticle();

    public ShoppingCart() { // Constructor }

    private void jbInit() throws Exception {
        /** @Label */
        shopping_cart_Label.setText("Shopping Cart");
        ...
        /** @Button */
        remove_article_Button.setText("Remove article");
        remove_article_Button.setEnabled(false);
        remove_article_Button.addActionListener(new java.awt.event.ActionListener() {
            public void actionPerformed(ActionEvent e) {
                remove_article_Button_ShoppingCart_GUI();
            }
        });
        ...
        /** @List */
        selected_articles_ScrollPane.getViewport().add(selected_articles_List, null);
        selected_articles_List.addListSelectionListener(...);
        /** @Panel */
        this.getContentPane().add(close_Button, null);
        ...
        /** Listener on children */
        confirmremovearticle_GUI.cancel_button.addActionListener(...);
        confirmremovearticle_GUI.accept_button.addActionListener(...);
    }

    /** @Button */
    void remove_article_Button_ShoppingCart_GUI() {
        this.setEnabled(false); confirmremovearticle_GUI.setVisible(true);
    }

    /** @Button */
    void close_Button_ShoppingCart_GUI() { this.setVisible(false); }

    /** @List */
    void selected_articles_List_ShoppingCart_GUI() {
        remove_article_Button.setEnabled(true);
    }

    /** @Methods children */
    void cancel_button_ConfirmRemoveArticle_GUI() { this.setEnabled(true); }

    /** @Button */
    void accept_button_ConfirmRemoveArticle_GUI() { this.setEnabled(true); }
}
```

This work was previously published in Visual Languages for Interactive Computing: Definitions and Formalizations, edited by F. Ferri, pp. 328-356, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 2.5
Design Methods for Experience Design

Marie Jefsioutine
University of Central England–Birmingham, UK

John Knight
University of Central England–Birmingham, UK

ABSTRACT

The following chapter describes an approach to Web design and evaluation where the user experience is central. It outlines the historical context in which experience design has evolved and describes the authors’ experience design framework (EDF). This is based on the principles of user-centred design (UCD) and draws on a variety of research methods and tools to facilitate the design, development, and evaluation of user experiences. It proposes that to design usable, accessible, engaging, and beneficial Web sites, effort needs to focus on visceral, behavioural, reflective, and social factors, while considering contexts such as the who and why; what and how; when and where; and with what of Web site use. Research methods from a variety of disciplines are used to support exploration, communication, empathy, and speculation. Examples of the application of the EDF, to various stages of the Web design process, are described.

INTRODUCTION

Although change is afoot, designers (including design engineers)—still worry that corporate bosses and clients see them as merely ‘making things pretty.’

— Molotch, 2003, p. 28

Producing Web sites is a process that involves a range of skills and disciplines. Design is not an add-on to make the screens look good. Design impacts on what people do, how they do it, and with whom they do it. Products, whether they are Web sites, toasters, or services embody a range of values and our interaction with them is a form of communication. When interacting with a Web site, for example, we may be communicating with other people, real or imagined. As we type with the keyboard, we hear and feel the keys and may vent our frustration through the force of our tapping. While reading the words on a screen we are also taking in all kinds of nonverbal messages...
from the layout—images, colours, fonts, icons, language, and style of language. Our experience is also affected by our surroundings, our memories of past actions, our current knowledge, and our expectations of the future.

Interaction with a Web site is mediated by a network of people, machines, and systems. The Internet works because of the design of protocols, browsers, laws, technical standards, security mechanisms, machines, communication technologies, and physical infrastructure as much as it does because of the design of the Web pages. Our experience of a Web site is therefore a product of our personal understanding of its context. Givechi and Velázquez (2004) describe the “positive space” of a product, which becomes more significant than the product itself, “the aura of a product, the sum of its physical attributes plus its intangible essence—or the meaning it hosts for each of its users” (p. 43). The meanings we attach to a product change with time, in part due to our changing experiences, the wider socioeconomic and political context of a product, and our changing expectations. To build a successful product, a designer needs to be aware of as many of these factors as possible, and this is why a consideration of the whole user experience is important.

FROM SOFTWARE ENGINEERING TO EXPERIENCE DESIGN

Concepts like “user experience” and “design of experience” are common in the design and business communities now (Fulton-Suri, 2004, p.14), but this has not always been the case, and in some software development circles designer still has negative connotations. It is important to recognise the context in which Web design has emerged, to understand why these concerns are so topical.

The development of computer software and interfaces in the 1960s emerged from fields traditionally associated with engineering and science. This era was typified by optimism for technology, and even in the traditional design disciplines there was a movement towards rationalising design methods. Hailed as the “design science decade” (Fuller, 1969, p. 305) the emphasis was on objectivity, rationalism, and technology.

The term software engineering can be traced back to an international conference convened by NATO in 1968, to solve the “software crisis” (Campbell-Kelly & Aspray, 1996, p. 200). The crisis emerged from large scale (often military) software projects that encountered problems in management and quality. In response to these failings a number of risk management strategies were advocated including the waterfall development lifecycle. Here design and development flowed through a predetermined course of phases. The formalised nature of the process was predicated by each phase having specified inputs and outputs, which could be checked along the way.

A design approach based purely on linear, logical problem solving did not work for the new technologies (e.g., computing, software, and solid state electronics), where problems were nebulous and constantly evolving through technological development and changing requirements. Brooks’ (1975) experience on IBM’s System/360 project was typical. He describes how software defied traditional logical engineering approaches. Even increasing resources did not improve the success of the approach.

In design disciplines such as product design and architecture, there was a growing realisation that some problems were unsolvable by logical deduction (Cross, 2001). Rittel and Webber (1973), for example, contrasted the “wicked problems” of design with the “tame” ones of science and engineering. In software design attention was turning to human factors. Whereas traditional engineering and industrial design focussed largely on external and measurable ergonomic factors, computer interaction required an understanding of internal, cognitive aspects. In 1969 the International Journal of Man Machine Studies was launched.
and a new field emerged, looking at what is now called human-computer interaction (HCI).

The idea that the computer interface could be modelled on human needs rather than driven by system capabilities led to the development of the graphical user interface (GUI) launched with one of the early commercially available personal computers, the Xerox Star, in 1981. Although the product itself was not successful, the concept of the GUI became almost universally adopted. This interface had been developed using usability engineering methods, including paper prototyping and testing on potential users. The design process was iterative and responsive.

In 1985, Gould and Lewis codified their own approach to software development, identifying the following key principles: early focus on users and tasks; empirical measurement through early and continual user testing; integrated design; and iterative design (Gould 1995; Gould & Lewis, 1985). UCD, as it came to be called, is now embodied in international standards for software development. International Organization for Standardization (ISO, 1999) 13407, for example, describes it as:

an approach to interactive system development that focuses specifically on making systems usable. Whatever the design process the allocation of responsibilities and roles adopted, the incorporation of a human-centred approach is characterised by the following:

a. the active involvement of users and a clear understanding of user and task requirements;

b. an appropriate allocation of function between users and technology;

c. the iteration of design solutions; and

d. multidisciplinary design.

Usability became a high profile issue in the 1990’s (see Knight & Jefsioutine, 2002) and much of the HCI literature aimed at supporting interaction designers, focused on cognitive and behavioural models of human interaction with attempts to formalise methods (e.g., Harrison & Thimbleby, 1990) and generate guidelines about human behaviour, much like the rational approaches to design.

The relationship between humans and computers is not entirely rational or logical however. Reeves and Nass (2002), for example, describe the ways in which people treat computers and software like real people with feelings and personalities. Jordan (2000) argues that usability is no longer sufficient as a design goal. He argues that customers now expect products to be easy to use and claims that “usability has moved from what marketing people call a ‘satisfier’ to being a ‘dissatisfier’” (Jordan, 2000, p. 3). People no longer notice when a product is usable, just when it is difficult to use. Fulton-Suri (2004) points out “established products have become more similar in technology, functionality, price and quality, companies have turned to design to differentiate, their offerings…to create stronger emotional connections with their customers” (p. 13). Lastly, the notion of a problem that can be identified and solved through logical deduction is often at odds with how successful products are developed and adopted by consumers.

The challenge for Web designers is no longer just to produce a functioning and usable product, they must now meet the needs of an increasingly sophisticated and demanding audience and a competitive market place. The EDF was developed by the authors to address these challenges by collating useful methods and approaches to designing products like Web sites and by considering the whole user experience (Jefsioutine & Knight, 2004).
THE EXPERIENCE DESIGN FRAMEWORK

The EDF advocates the principles of UCD already described, and adds another set of considerations: qualities, dimensions of experience, research contexts, and research methods. These are described hereafter.

Qualities

Each product will have its own specific user requirements. The EDF proposes that four fundamental qualities underlie these requirements.

Usable

To go beyond usability does not mean that it is no longer necessary. The benefits of usability are well documented (e.g., Bevan, 2000). Usability is defined as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” (ISO, 1998) and “the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions” (ISO, 2000). Indeed, in long-term use of a product, there is evidence that usability becomes more important than style in predicting users’ satisfaction (Maguire, 2004). Usability cannot be ignored as a design goal.

Accessible

The Web relies on a level of standardisation such that pages can be accessed by anyone, on whatever browser or device they use, be it a PC, Mac, or mobile phone. As the Web has grown, so have the standards and so have the number of inaccessible Web sites. Despite the publication of guidelines for accessible web design by the World Wide Web Consortium (W3C, 1999), a survey of Web designers found that difficulty interpreting guidelines was a major barrier to implementing them (Knight & Jefsioutine, 2003). Furthermore, the UK’s Disability Rights Commission (2004) found that nearly half of the usability and accessibility problems were not violations of any of the WCAG’s checkpoints. DiBlas, Paolini, Speroni, and Capodieci (2004) argue that “W3C guidelines are not sufficient to ensure an efficient—even less satisfactory—Web experience” (p. 89). It is important that accessibility is seen as part of the user experience rather than a series of technical checkpoints to cover.

Figure 1. The experience design framework (Jefsioutine & Knight, 2004)
Engaging

Shedroff (2001, p. 4) suggests an experience comprises of “an attraction, an engagement, and a conclusion”. What attracts someone to a product could be a need to perform a task, an aesthetic quality, or an affordance. The engagement is then sustained over a period of time, beyond the initial attraction. Csikszentmihalyi (1991) describes the experience of optimal experience and flow: “Concentration is so intense that there is no attention left over to think about anything irrelevant, or to worry about problems. Self-consciousness disappears, and the sense of timing becomes distorted” (p. 71).

Although this level of engagement may not be appropriate for all products, it is useful to consider the properties of an experience that make it engaging. Fiore (2003) suggests that an experience includes a number of dimensions: it is educative and memorable; whole, unique and nonreproducible; historical; meaningful/aesthetic; contextual; physical/sensual/embodied; and situated in time and space. Jones, Valdez, Nowakowski, and Rasmussen (1994) suggest that engaged learning tasks are challenging, authentic, and multidisciplinary. Such tasks are typically complex and involve sustained amounts of time... and are authentic. Quinn (1997) suggests that engagement in learning applications comes from two factors—interactivity and embeddedness, where the user perceives that they have some control over the system, and it is relevant and meaningful to them.

Beneficial

According to Csikszentmihalyi (1991), an optimal experience is so gratifying that “people are willing to do it for its own sake, with little concern for what they will get out of it” (p. 71). One might assume that what they are getting is some degree of pleasure. Jordan (2000) identifies pleasure as the ultimate quality of the user experience, and DeJean (2002) points out that pleasure is a complex concept. Apparently unpleasant aspects of the user experience, such as difficulty, challenge, and fatigue can all be pleasurable in certain contexts, for example, by generating feelings of achievement or superiority.

Bonapace (2002) adapts Maslow’s (1970) hierarchy of human needs to product use qualities. His pyramid begins with safety and well-being, moves up to functionality, and then usability which leads up to an apex of pleasure as the ultimate quality. Dunne (1999) explores the “aesthetics of use” and argues, “The most difficult challenge for designers of electronic objects now lies not in technical and semiotic functionality, where optimal levels of performance are already attainable, but in the realms of metaphysics, poetry and aesthetics, where little research has been carried out” (p. 7).

Liu (2003) describes qualities of “psychosomatic soundness,” referring to the extent that a product contributes to the “wholesomeness” or well-being of a person (from “harmful” to “healthful”) and the degree to which it is ethical (“bad/wrong” to “good/right”) (p. 1296). Knight (2004) too, argues for a reconsideration of the ethics of HCI design. The EDF uses the term benefit to include such concepts as ethics, psychosomatic soundness, pleasure, and self-actualisation.

The EDF, therefore, advocates designing for the four fundamental qualities of product use of accessibility, usability, engagability, and benefit. To do this it becomes necessary to widen the focus of research beyond cognitive and behavioural interaction, typical of usability, and HCI studies, to include a multi-dimensional approach to experience.

Dimensions of Experiencing

McDonagh-Philp and Lebbon (2000) suggest that emphasis must change “from hard functionality, to soft functionality” (p. 38). Rather than focusing on what a product does and how it does it, the focus is on less tangible aspects like emotional
associations, familiarity, aesthetics and taste. Fiore’s (2003) framework of experience considers physical, emotional, and intellectual aspects. In a similar vein, Norman (2004) describes emotional design in terms of its “visceral, behavioural, and reflective” elements (p. 63). He points out that these three levels of experience are not discrete but interact with each other. Spillers (2004) suggests, for example, that a new icon on a screen could arouse a state of curiosity or annoyance, producing a change in the user’s emotional state “which can either propel the user toward a feeling of satisfaction (success) or disappointment (failure)” (p. 2).

Jordan (2000) develops Tiger’s (1992) concept of pleasure, and describes four ways in which pleasure can be experienced: “Socio-Pleasure” arises from interaction with others or from a product that represents a social grouping; “Psycho-Pleasure” comes from the satisfaction felt when a task is successfully completed or from a product that makes a task more pleasurable. “Physio-Pleasure” is derived from the senses; and “Ideo-Pleasure” is derived from entities such as books, art and music or the values that a product embodies (p. 13-14).

The EDF uses Norman’s (2004) classification and adds a social dimension. The EDF, therefore, directs attention to the visceral, behavioural, reflective, and social dimensions of experience. Furthermore, it suggests that they be considered in the context of the use qualities. For example, considering accessibility in the context of each dimension may generate design goals such as:

- **Accessible/visceral**: legibility and visual clarity, text alternatives to audio content
- **Accessible/reflective**: limiting cognitive overload / clear, simple language use
- **Accessible/behavioural**: keyboard alternatives to using a mouse, shortcuts, voice input
- **Accessible/social**: culturally inclusive, secure, private, moderated

### Research Contexts

This section describes a set of key questions that can be asked throughout the design process to develop an understanding of the contexts of product use. The questions are derived from a number of models including Rothstein’s (2002) model consisting of the four As of “activity, artefacts, atmosphere and actors” (p. 3), and Ortony, Clore, and Collins (1998) cognitive model comprising “events, agents and objects (p. 63). The EDF advocates four key contexts: (1) who and why (users and stakeholders and their motivations); (2) what and how (content, tasks, task flow, actions, functionality); (3) when and where (situation, frequency, environment); and (4) with what (tools, knowledge, and skills).

### Who and Why

In order to be user centred it is necessary to identify who the users and stakeholders are and what motivates them to use a product. It is important to include everyone, however limited their involvement with a product might be. Users of a software system, for example, might include people that buy it, use it at home or at work, communicate through it, people who sell it, or sell on it, administrate it, repair it, install it or support it (e.g., Hackos & Redish, 1998).

When the focus of UCD goes beyond usability, it becomes necessary to collect data that pertain to the qualities and dimensions of the EDF. This might include: demographics (such as age, gender, ethnic origin, and culture); behaviour and skills (such as computer literacy, typing skills, embedded knowledge of a task or system); knowledge and experience (such as novice or domain expert, tacit knowledge); personal characteristics and motivations (such as personality, learning style, attitude, aspirations, values, beliefs, tastes, and preferences); and physical characteristics (such as dexterity, physical abilities, height).
What and How

This refers to the tasks or activities that will be supported, influenced, or affected by the product and how users carry out these tasks. Tasks are typically mapped from a behavioural perspective or cognitive dimension. Liddle (1996) suggests, “The most important component to design properly is... the user’s conceptual model. Everything else should be subordinated to making that model clear, obvious, and substantial” (p. 21). The EDF suggests that activities be considered in all dimensions. For example, an airhostess may be performing complex emotional tasks as well as checking in baggage (customer relations, anxiety reduction, risk assessment, etc.) (Hochschild, 1983).

When and Where

The role of context in understanding tasks was emphasised by Suchman’s (1987) notion of situated actions, that activity is conditional on the situation in which it takes place and is of an improvised rather than planned nature. What people say they are doing or how they do it, is an after-the-event rationalisation. Context should be considered in terms of all of the dimensions. For example, what is the user’s current frame of mind? What ideological factors will influence their experience? What aspects of the environment affect their visceral experience—are there loud distracting noises around or reflections on the screen? What emotional state is the user in? What other behaviours are they performing? What is the social context—is the product being used with friends to communicate, to play, to learn, or to work cooperatively?

With What

This refers to objects, artefacts, or tools that are being used or are influencing use (such as software, browsers, input devices, assistive technologies), and to knowledge, expertise, and skills that are used to carry out tasks. Considered in the context of qualities and dimensions, brainstorming, for example, might suggest that a haptic interface or 3D glasses (visceral) may improve engagability when viewing a virtual museum object. Conversely, users’ knowledge of existing interface conventions (reflective) may create expectations that reduce the usability of an innovative interface.

Methods and Tools

Reeves and Nass (2002) point out that people are rarely aware of their less rational motivations, so attempts to model what people are thinking or doing by asking them will not necessarily capture the reality of an experience. They advocate the use of methods from the social sciences to establish what people really think. The EDF suggests casting a wide net across many disciplines to find appropriate and useful methods.

Rather than prescribing a process or method, the EDF suggests that a range of tools and techniques can be employed provided they cover four basic purposes—exploration, empathy, communication, and evaluation. Furthermore, by applying these methods to the dimensions, qualities, and research perspectives, a better understanding of the user experience as a whole can be achieved.

Exploration

These methods are about discovery and can be drawn from demography, ethnography, market research, psychology, and HCI. They include surveys, interviews, questionnaires, focus groups, task analysis, field observation, user testing, affinity diagramming, laddering, and experience diaries. A key area of exploration is in understanding users’ mental models of a domain. This can be explored by, for example, in depth elicitation techniques and card sorts.

Contextual interviews are conducted during the activity or in the environment in which the product will be used. Users are able to refer
Design Methods for Experience Design

to artefacts, such as the documents, memos, and equipment that they normally use in their workflow, and indeed may be encouraged by the researcher to describe what they are doing. One of the most common methods of achieving this is to use the “think aloud protocol” (where users verbalise their actions). This has the advantage that the user is not relying on memory to describe his or her actions, and the researcher is able to note and probe omissions. There are some disadvantages however. The act of conscious reflection may change the way the task is performed and may create excessive cognitive load, or compete for the same cognitive channel as the task.

Both ethnography and ethnomethodology have been applied to eliciting contextual data. Ethnography is the study of a culture achieved by researchers immersing themselves in that culture in order to understand it, while ethnomethodology studies the ways in which participants give order to and make sense of their social worlds (Garfinkel, 1967). The focus of ethnomethodological studies is often at a very detailed level of interaction, including interaction between people and artefacts, technologies, and systems. For this reason it is a useful approach for the study of complex work situations, for example, air traffic or train control systems (Heath & Luff, 2000). The researcher may also use the artefacts produced by organisations, such as manuals and training materials; policies and procedures; and forms and documentation to enrich the detail of the observation. The rationale for applying these approaches to design is that the knowledge acquired will enable designers to work with, rather than against, users’ ways of understanding and making sense of the activities in which they are engaged.

Nevertheless, integrating these methodologies within the design process is not without its problems. The skills of trained ethnographers are often underestimated. Done properly, ethnographic methods involve sophisticated and skilful observation and recording techniques. The data may be far more detailed than the designer needs, and in a form that is difficult to interpret. Furthermore it can take many years to do a full ethnographic study, and for most software development projects, this amount of time is unrealistic. A simpler version may include shadowing target users for a short time, noting what they do and use, taking photographs, asking questions, and using techniques to elicit their values or emotional responses (Beyer & Holtzblatt, 1998).

Communicating

Design involves communication with a wide range of people, from users to software engineers. Design teams need to accommodate different viewpoints and share a common language. Curtis (2002) emphasises the importance of actively listening to a client and finding out the story behind a product. Hammer and Reymen (2004) stress the importance of designers expressing their emotional as well as rational reflections on design decisions. Communication methods serve to clarify and share the goals of stakeholders, the exploratory research data, design requirements, and ideas to a multi-disciplinary team who may not have a common vocabulary. It is important to ensure good communication throughout the process to ensure the product itself communicates the design goals effectively. Methods include story telling; user profiles and personas; use cases or task scenarios; scenario-based design; mood boards; written briefs and specifications; storyboarding; and prototypes.

User profiles are generated from demographic data and lifestyle surveys. They can include textual and visual descriptions of key user groups. They are used to think through design solutions and for recruiting users for research. Mood boards are normally collages of photographic information that aim to generate a visual “personality” for the product or service. Mood boards can be created by designers or users or be the result of collaboration between the two. Mood boards are useful because they work at a nonverbal level where
people may otherwise have difficulty expressing their wants and needs. Storyboards are time-based often with a narrative aspect. Storyboards can be sketches or low-fidelity screen shots of a user’s interaction. Prototypes range from paper sketches to working interactive replicas. Information architecture can be communicated through formalised diagrams and charts or simple tree structures and hierarchies.

Empathy

These methods represent an approach aimed at gaining a deeper understanding and empathy for users. They include focus groups, diaries, workshops, participatory design, and immersion. Diaries can be used to record users’ interaction with a product over time while workshops and participatory design involve users in the development team either directly or through a user advocate that champions their perspective. Participant observation, or “eat your own dog food,” involves taking part in the activity or culture being observed, where the designer becomes a user. Molotch (2003) describes a method used by Ford designers in which they test products “dressed in what they call a ‘third age’ suit, with glasses and gloves, to simulate having the body and eyesight of a 70-year old” (p. 49).

Crossley (2004) describes a technique used to help designers develop an awareness of the target audience for male grooming products. It involved:

*rapidly immersing the design team into the lives, hearts and minds of people in a short space of time. The challenge for this project was to get young men inspired to tell us their own stories and express their emotions about a mundane functional activity*”… “Character modelling [was used] … where the team and sometimes the user has a kit with questions, cameras and collages, [that enabled them] to frame and understand the lifestyle of the person they are creating for.” (pp. 38-39)

Personas are used to develop a shared vision of end users among development teams. They can be textual, visual, animated, or acted. One of the most important functions of personas is to get teams to think differently about people, as Cooper (1999) notes:

A fully realized, thoroughly defined user persona is a powerful tool. Until the user is precisely defined, the programmer can always imagine that he is the user. A completely pronounced user persona is key to the suppression of any tendency for the developer to usurp or distort the user persona’s role. Long before a single line of code is written, a well-defined user persona becomes a remarkably effective tool for interaction design. (pp. 128-129)

Role-play is often used to encourage designers to empathise with users. Dramatic techniques can also help teams get into the minds of users and their values. Carmichael, Newell, Morgan, Dickinson, and Mival (2005) describe how the UTOPIA project improved designers’ understanding of requirements. Using professional actors the project developed video scenarios of elderly people using technology. This was then shown to design teams to encourage them to design for others rather than themselves.

Speculation

In addition to understanding users’ wants and needs, designers also need to speculate about new solutions and future trends. The Sony Walkman, for example, introduced an entirely novel mode of behaviour that no users had asked for. The decision to add text messaging to mobile phones was based on a speculation of how that functionality might be needed and by whom. The success of text messaging, however, was based on its uptake by an entirely different user group with its own needs and method of use. Designers may need to predict how users will adopt and adapt to a new product.
Design Methods for Experience Design

Here the solution is not tied to a particular technology or need but to reconceiving it. Good design requires up-to-date knowledge of what is possible and the ability to see beyond the obvious. There are a number of methods that achieve this. Jones (1990) sees speculation as “divergence, transformation and convergence” (p. 64-68) and suggests methods such as brainstorming (search for alternatives) and synergetics (search for alternatives by analogy).

Rosson and Carroll (2002) describe scenario-based design in which scenarios are deployed throughout the design process to speculate the future. Scenarios can be used to challenge existing practices and encourage users and designers to think beyond the confines of the current situation. Dunne (1999) describes the use of scenarios to communicate design ideas and to stimulate thinking beyond the preexistent reality. He suggests that they can “push the viewer towards a more complex, emotional or revolutionary understanding of the problems posed” (p. 75). By testing design ideas rather than prototypes, the user’s attention is shifted from the “aesthetics of construction” to the “aesthetics of use” (p. 73).

Speculation and innovation are, in part, about predicting the future, and a number of methods have been developed for this. The UK Cabinet Office published a Futurist’s toolkit (UK Cabinet Office, 2001) describing six key methodologies for futures work, including quantitative and qualitative trend analyses, the Delphi survey method, scenario methods, wildcards, and future workshops. Cayol and Bonhoure (2004) describe a methodology which used sociological studies predicting the evolution of French people for the year 2025, to identify and evaluate future product concepts.

De Bono (1995) uses lateral thinking to generate different scenarios and ways of thinking about problems. De Bono’s “Six Thinking Hats” (pp. 78-79) is a method for creative thinking but is also useful for developing empathetic intelligence. This method involves six team members role-playing different aspects of the design process. Thus the white hat is neutral and reflects back opinions and ideas, whereas the black hat is judgemental, cautionary and avoids mistakes. Speculative methods are important because design teams need to be able to see beyond user requirements and consider latent needs and potential opportunities.

Evaluation

Evaluation methods include auditing, standards compliance, and user testing. Evaluating may also use similar methods to exploration, with a shift in emphasis from discovery to checking outcomes against intentions. Does the product meet the design goals and/or the user expectations? Evaluation can be formative, conducted during the development of a product, or summative, conducted when a product is complete and is being used. Summative testing of an earlier version or a similar product can be useful to identify design goals, while summative testing of the product at the end of its design lifecycle is usually done for auditing and verification purposes. Feedback at this stage is of little use to the designer, the deadline has passed and the money is spent. Clearly formative testing is most helpful to a designer/developer. Gould and Lewis (1985) stress the importance of empirically testing design iterations throughout the design process. Evaluative tools such as heuristics are often used, although evidence suggests that they are no substitute for testing real users (e.g., Lee, Whalen, McEwen, & Latremouille, 1984). The EDF broadens the test and evaluative criteria from the traditional focus on cognitive and behavioural measures. Bonapace (2002) describes a method aimed at tapping into the four pleasures described by Jordan (2000), called the Sensorial Quality Assessment Method (SEQUAM) applied to the design of physical products in car manufacturing. User testing can combine empirical methods of behavioural observation with techniques such as co-discovery, think aloud and empathic interviewing, to tap into
the subjective aspects of experience.

Establishing evaluation criteria at the begin-
ning of a design process helps to focus the evalua-
tion process from the beginning, although it is wise
to allow some freedom for the evolution of design
goals through iteration and to allow evaluative
concerns to emerge (Hall 2005). The EDF can be
used to generate evaluative criteria for qualities
in each dimension and context of use.

**APPLYING THE EXPERIENCE DESIGN FRAMEWORK**

The authors have used the EDF to adapt and focus
methods for requirements research, brief develop-
ment, ideation, and testing and have developed
a range of services and training based on it. The
EDF has been particularly useful in generating
design goals, aiding decision making, and devel-
op ing user-testing scenarios. Some examples of
applications follow.

**Visioning Workshops**

Visioning workshops usually take place at the
beginning of the design process, preferably before
the brief is finalised. The workshop is structured
around the experience design framework and ful-
fil s a number of functions. Techniques are adapted
to improve communication; to build literacy and
a common language to describe the medium; and
to build an empathic understanding of users and
other contexts of use. Individual and group activi-
ties are facilitated by a researcher and recorded
by a scribe. Activities include the following:

- Participants identify their vision of the
  project goals and their personal objectives,
  roles, and stories.
- Participants share their personal preferences
  and tastes by discussing examples of the
  product type or medium from visceral, be-
  havioural, reflective, and social perspectives,
  and in so doing, build up a shared language
  for product qualities.
- Participants discuss the nature of the design
  problem from different perspectives and
  use speculative techniques to generate new
  ideas.
- Participants identify audience and stakehold-
  ers; then develop personas and scenarios
  based on different contexts of use.
- Participants share desired product qualities,
  functionality, and content from the perspec-
  tive of the stakeholders identified.
- Participants identify and prioritise product
goals and develop evaluative criteria.

By the end of the workshop, participants have
prepared the foundations of a creative brief, have
started working together as a team, and have
formed a common language and shared under-
standing of the project goals.

**Contextual Interviews**

Contextual interviews take place prior to detailed
requirements specifications and may follow a
visioning workshop. Firstly, stakeholders are
identified and a representative sample recruited.
The aim is to survey a large sample and to it-
teratively develop knowledge about the design
problem. In this context, as many as 50 users may
be involved and the focus is to gain as full a set
of requirements as is possible in a short space of
time. The interviews are semi-structured over
approximately 30 minutes and are conducted
within the context of use. Their exact format is
dependent on whether the product is a new one or
a refinement of an existing one. In the former case,
the researcher works with low-fidelity prototypes
and in the latter case with the existing product.
Activities include the following:

- Stakeholders are asked to provide docu-
  ments, processes, and artefacts involved in
  the tasks.
Interviewees are asked to complete a persona template with details about themselves, their interests, and their lives.

- Interviewees are asked to identify critical tasks, events, and work-arounds with the existing product.
- Interviewees identify relationships with other users/stakeholders in their use of the product, and what other tools or products they use in association with the tasks.
- Interviewees are asked to describe key tasks with the product and/or walk through a task. The interviewer elicits details of any prior knowledge, expertise, or skills being applied to the tasks and probes for emotional responses to aspects of the activity.
- Interviewees are asked to describe how the product fits in with their daily life.

The results of interviews inform the production of anonymous personas and use scenarios, which are used to communicate the requirements to the development team and build their empathy with the users.

**Conceptual Design Workshops**

Conceptual design workshops involve development teams rethinking the design problem and considering potential solutions. The aim is to generate a number of alternative design concepts. These are then evaluated and then “worked up” for initial user testing or participatory prototyping (see what follows). Workshop members usually include members of the development team plus representatives of key stakeholders. The workshops are structured around the EDF and activities include the following:

- Participants reconceptualise the design problem in terms of the EDF’s contexts.
- Participants brainstorm and develop a number of design concepts.
- The concepts are discussed and similar solutions are merged together and evaluated against the EDF’s qualities.
- The process is repeated until a number of distinctly different and viable concepts have been generated and can be tested on users.

**Participatory Prototyping**

Participatory prototyping combines the skills of the development team with user feedback. Prototypes are developed of content structures, interaction flow, and layouts. Initial prototypes are developed and users are asked to critique or adapt the prototype with specific reference to the qualities of the EDF and the dimensions of experience. Designers interpret this feedback and develop further prototypes, with the focus on interaction and structure, and later on look and feel issues. Activities include the following:

- Users are asked to card sort content or to map out their expectations of content domains and structures.
- Users are asked to “walk through” prototypes in order to carry out their key tasks, using the think aloud method.
- Responses to the organisation of the interface and the terminology used are elicited.
- Users perform tasks with the prototype and are asked to speculate on improvements, applications, or additional features.
- Where trade-off or multiple interfaces exists users sort the prototypes on the basis of their preference.

By the end of the process a complete low-fidelity prototype has been developed that has been iterated around the qualities of the EDF.

**Audience Reception Workshops**

Audience workshops review the final prototype design. Users are recruited to represent the key
stakeholder groups. The prototype then undergoes a group critique that tests the solution against the initial requirements and evaluative criteria gathered by the visioning workshop and contextual interviews. As well as ensuring that the prototype is suitable the workshops gauge barriers and opportunities to take up and adoption of a new product or design. In addition, it provides the development team with a rationale and evidence of the suitability of the final design solution. Activities include the following:

- Group walk-throughs of the prototype
- Identification of conflicts and trade-offs
- Comparison of look and feel prototypes
- Quantitative research methods and user attitude measurement

**Post Implementation Research**

Post implementation research reviews the application after launch. Unlike traditional usability that focuses on requirements and development, this research is necessary to monitor changing user needs throughout the product’s lifecycle. A user cohort is recruited and is asked to provide regular feedback based on the EDF. As well as ensuring that maintenance and new features are accepted by users this research is important to identify new products and user needs. Activities include the following:

- User diaries
- Online discussion groups and surveys
- Focus groups and ongoing user testing

**CONCLUSIONS AND THE FUTURE OF WEB EXPERIENCES**

The EDF was created in response to the needs of clients and has generated research, which has fed back into the design of services. The EDF has been applied to a number of internal and external products and has been refined through practice. The EDF is meant to provoke discussion, raise questions, challenge assumptions, and generate alternatives.

Every design problem is different and so it is necessary to deploy a range of research methods to support the work of the digital media designer. Nevertheless, these methods should be focused around the key issues for UCD, which is to understand users, the tasks they undertake, and the contexts in which they function. They provide a rich understanding of users from which the creative designer can create usable and desirable products. The EDF helps teams focus requirements research and can be used in a workshop environment with a range of stakeholders and users. The EDF can be used as a brainstorming tool that aims to map out the requirement and the process, and possible solutions from the team and stakeholders.

If methods are focused around qualities, dimensions, and contexts of experience, they provide a richer understanding of users from which the designer can create usable and desirable products. The EDF provides a reminder of the complexity of the user experience and can be used throughout the design lifecycle.

**REFERENCES**


Cooper, A. (1999). The inmates are running the asylum. Why high-tech products drive us crazy and how to restore the sanity. Indianapolis, IN: SAMS.


*This work was previously published in Human Computer Interaction Research in Web Design and Evaluation, edited by P. Zaphiris, pp. 130-147, copyright 2007 by Information Science Publishing (an imprint of IGI Global).*
Chapter 2.6
Integrating Usability, Semiotic, and Software Engineering into a Method for Evaluating User Interfaces

Kenia Sousa
University of Fortaleza, Brazil

Albert Schilling
University of Fortaleza, Brazil

Elizabeth Furtado
University of Fortaleza, Brazil

Abstract
We present artifacts and techniques used for user interface (UI) design and evaluation, performed by professionals from the human-computer interaction (HCI) area of study, covering usability engineering and semiotic engineering, which can assist software engineering (SE) to perform usability tests starting earlier in the process. Tests of various interaction alternatives, produced from these artifacts, are useful to verify if these alternatives are in accordance with users’ preferences and constraints, and usability patterns, and can enhance the probability of achieving a more usable and reliable product.

Introduction
In a software development process (SDP), it is crucial for developers, customers, and users to interact in order to specify, generate, and evaluate the software. From software specification to its delivery, various kinds of tests must be performed, involving aspects such as: functionality, portability, performance, and usability. This work focuses on the context of usability, communicability, and functionality tests (e.g., appropriateness of a chosen interface design alternative to user preferences, consistency to a visual pattern, efficient execution of interactive tasks on interface objects, etc.).
Through our researches on tests in HCI and SE, and through our experiments on their integration in a SDP, we verified that HCI concepts facilitate the UI evaluation work performed by the test team of an interactive system under development. More specifically, by means of UI generation based on HCI models (e.g., task model), it is possible to evaluate the UI earlier (e.g., its functionality), independent of having the entire noninteractive specification ready. Prototypes, for instance, can represent UI design alternatives that may be tested early by HCI experts to verify if they are in accordance with user preferences, usability patterns, and so on.

This work presents a SDP to design and evaluate UIs, based on the integration of concepts, models, and activities of usability, semiotic, and software engineering.

This chapter is structured as follows: The “User-Interface Evaluation” section shows the contribution of each engineering area to UI evaluation; “The Process” section describes the UI design process; “The Evaluation Strategy” section describes the UI evaluation process, showing which concepts are used to perform tests and when they are performed; the “Case Study” section describes the case study in which we designed and evaluated UIs for the Brazilian System for the Digital Television (SBTVD); and, finally, the “Findings and Future Works” section describes findings and future works, and the “Conclusion” section concludes this work.

**USER-INTERFACE EVALUATION**

In this section, we present concepts and evaluation techniques from usability engineering, software engineering, and semiotic engineering.

**Usability Engineering**

Usability engineering is a set of activities that ideally take place throughout the lifecycle of the product, with significant activity at the early stages even before the UI has been designed. The need to have multiple usability engineering stages supplementing each other was recognized early in the field, though not always followed in development projects (Gould & Lewis, 1985).

In usability engineering, techniques and methods are defined aiming to assure a high usability level of the interactive UIs. Among them, we emphasize the application of ergonomic criteria in the UI design. Verification of these criteria in designed UIs is called heuristic evaluation, performed by usability experts without user participation. Evaluators examine the IS searching for problems that violate general principles of good UI design, diagnosing problems, obstacles or barriers that users will probably encounter during their interaction. In addition, methods to capture usability requirements attend to user preferences, restrictions, and use-context. A usability requirement can be derived from an interaction restriction; such as if part of the system needs to be implemented for palm-top devices.

The evaluation approaches from usability engineering suggests a structured sequence of evaluations based on “usability inspections methods” and on “usability tests”.

Some inspection methods are: (1) heuristic evaluation, verification of usability heuristics (Nielsen, 1993); (2) review of guidelines, verification if the UI is according to a list of usability guidelines (Baranauskas & Rocha, 2003); (3) consistency inspection, verification of the consistency among the UIs related to terminology, color, layout, input and output format, and so on; and (4) cognitive walkthrough, simulation of the user “walking” through the UI to execute typical tasks.

Some usability test methods are: (1) thinking out loud, we request the user to verbalize everything he or she thinks while using the system, and we expect that their thoughts demonstrate how the user interprets each UI item (Lewis, 1982); and (2) performance measures, quantification of some evaluated items to make future comparisons.
Software Engineering

Software engineering is composed of technologies and practices that are used in the development of software products, enhancing software productivity and quality, by providing a more systematic and controlled SDP (Sommerville, 2001).

In software engineering, there are various types of tests to be performed in each test stage, such as: usability tests to ensure that access and navigation through functionalities are appropriate for users; UI tests, to ensure a good functionality of the UI components and verify conformity to corporate patterns; and functionality tests, which are responsible for verifying if the generated software achieves all the proposed functionalities according to the customer’s requests. Test cases (Myers, 2004) are normally generated and comprise procedures to be followed in test activities in order to deal with all possible situations when using the software, including basic flows, as well as error treatment and invalid data verification.

According to Pressman (1995), the main goal of test cases is to derive a set of tests that will probably reveal errors in the software. To achieve this goal, software engineering basically proposes two test categories: white-box and black-box tests.

For Pressman (1995), the white-box test must verify the internal part of the product, tests can be performed to guarantee that the components are integrated, and the internal operation achieves the performance level as specified in the requirements.

Functional tests, or black box tests, represent a test approach in which tests are derived from specifications of the system. In this kind of test, the evaluator is concerned with the functionality, not with the software implementation (Sommerville, 2001).

In these two test categories, software engineering defines four main types of tests: unit tests that are generally white-box tests; acceptance and regression tests that are usually black-box tests, and integration tests that blend this two categories.

Semiotic Engineering

Semiotic engineering is an HCI theory that emphasizes aspects related to the metacommunication designer user(s) via user-system communication, which passes through the UIs of interactive applications. The system is considered to be the “deputy” or a representative of the system designer (Souza, Barbosa, & Silva, 2001). The content of messages is the application usability model. Its expression is formed by the set of all interaction messages sent through the UI during the interaction process. The user plays a double role: interacting with the system and interpreting messages sent by the designer.

Semiotic engineering is essentially involved in test procedures with final users (empiric evaluation), aiming at system communicability analysis — based on qualitative evaluation, in which there are four phases: test preparation; labeling; interpretation, and formatting — and elaboration of the semiotic profile of the application to be evaluated. The techniques used in these phases are: system-user observations, questionnaires, (somative-evaluation) inspections, interviews, filming, and so on.

Semiotic engineering is essentially present in tests with final users (e.g., empiric evaluation), aiming at analyzing the system communicability. A UI has a good level of communicability when it is able to successfully transmit the designer message to the user, allowing him or her to understand the system goal, the advantages of using it, how it works, and the basic UI interaction principles.

Evaluation

After studying about evaluation techniques, artifacts and approaches from software, usability, and semiotic engineering we are able to conclude that an evaluation process can be seen under various perspectives.

Concerning software engineering, we noticed the importance of software quality concerning
functionality, performance, portability, and other nonfunctional requirements. Its artifacts and techniques include the evaluation of these aspects in an objective manner.

Usability engineering focuses in providing more ease of use, ease of learning, and efficiency to interactive systems.

Semiotic engineering includes procedures that allow the evaluation of the quality of the interactivity of systems by observing the communication through messages of the user to the system.

Based on these perspectives, we believe that an approach for UI evaluation of interactive systems that integrates these approaches is able to guarantee a system with quality concerning functionality, usability, and interactivity, derived from software, usability, and semiotic engineering, respectively.

Next, we will describe a lightweight development process for interactive systems, called UPi (Sousa & Furtado, 2004), which integrates HCI and SE activities, artifacts, and professionals.

THE PROCESS

UPi can serve as a guide, providing useful steps and artifacts that can be tailored and customized when organizations intend to develop usable interactive systems. One of the best advantages of UPi is the idea to focus on activities, artifacts and guidelines that add value to the UI generation. With this approach, it can be integrated with any other process and inherit activities that are vital to the entire process, but that are better defined and solidified in other processes. For instance, project management, configuration and change management, implementation, and deployment activities are very well detailed in the RUP (Kruchten, Ahlqvist, & Bylund, 2001). Besides the RUP, UPi can also be applied in conjunction with ISO 13407 (ISO 13407, 1999), which already has other activities defined and validated, such as project planning, testing, and so on.

UPi is composed of activities that aim at designing UIs. These activities are based on RUP activities, but they follow different guidelines that take into consideration usability aspects.

In this work, we are integrating UPi with UPi-Test (to be presented in the next section) in order to guide professionals that are developing interactive systems to evaluate them throughout the entire development process.

Phase I: Inception

The main goal in this phase is to elicit requirements from users in order to develop an interactive system that best suits their needs through the execution of some activities (presented as follows). These requirements are documented through certain artifacts: use-case models, task models, usability requirements, and paper sketches.

*Use-case models* represent a well-established manner to define the system functionality, while *task models* can be used to detail use cases by breaking them down into tasks. *Usability requirements* represent users’ preferences or constraints that can be part of a usable interactive system. *Paper sketches* focus on the interaction, UI components, and on the overall system structure, keeping the style guide secondary, without being too abstract.

The purpose of the *Elicit Stakeholder Needs* activity is to understand users, their personal characteristics, and information on the environment where they are located that have a direct influence on the system definition, and to collect special nonfunctional requirements that the system must fulfill, such as performance, cost, and device requests.

The purpose of the *Find Actors and Use Cases* and *Structure the Use-case Model* activities is to define the actors (users or other systems) that will interact with the system and the functionality of the system that directly attend to users’ needs and support the execution of their work productively.
The purpose of the Detail a Use Case activity is to describe the use case’s tasks using the task model, to describe any usability requirements related to the use case, to define the system navigation based on the task model hierarchical structure, and to create paper sketches.

The purpose of the Review Requirements activity is to verify, with usability experts, if the paper sketches are in accordance to the task model and validate, with users, if the requirements are in conformance with their needs by showing them the elaborated paper sketches.

**Phase II: Elaboration**

The main goal in this phase is to transform the requirements in a representation that can be understood by UI designers and programmers. These representations are provided by the following artifacts: system architecture, UI Definition Plan, and drawing prototypes.

System Architecture is composed of smaller components that represent the main functionality of the entire system. The UI Definition Plan is a new artifact that aims to define which visual objects should be part of the UI. Drawing prototypes produce an accurate image of the system and they are useful to demonstrate patterns and style guides.

The purpose of the Define and Refine the Architecture activity is to (re)design the classes that represent the data that are handled by users while performing certain tasks.

The purpose of the Define UI Plan activity is to define which visual objects and which usability patterns can be part of the UI according to the nonfunctional requirements defined in the Elicit Stakeholder Needs activity.

The purpose of the UI Prototyping activity is to design a UI prototype in drawings following the description specified in the task models, in the UI definition plan and in the system architecture.

The purpose of the Evaluate Prototype activity is to verify if the UI prototypes are in accordance to usability principles and to validate with users if the UI prototypes are in conformance with their needs.

**Phase III: Construction**

The main goal of this phase is to implement and verify the accuracy of the components implemented and of the UI designed.

The purpose of the Implement Components activity is to develop the classes previously designed and implement the UI prototyped.

The purpose of the Evaluate the Version of the System activity is to verify if the functionality of the interactive system is in accordance with users’ requirements.

**Phase IV: Transition**

The main goal of this phase is to deliver to the customer a system with high level of quality and usability.

The purpose of the Deploy the System activity is to make the system available for the customer.

The purpose of the Evaluate the System activity is to validate with users (by using the system in the deployment site) if the system conforms with their view of the system.

All of these phases are supported by processes concerning configuration and change management and project management, such as the RUP. This support is provided for the Manage Change Requests activity, which aims at evaluating the impact of change requests, deciding if they are to be included in the current iteration, and, if they are accepted, manage the changes in the appropriate artifacts.

Concerning the evaluation activities performed in each phase, they will be more thoroughly explained in the next section.


THE EVALUATION STRATEGY

The unified process for evaluating interactive systems “UPi-Test” (Schilling et al., 2005) has the same phases as the RUP (inception, elaboration, construction, and transition) (Kruchten, Ahlqvist, & Bylund, 2001). This process is based on the Unified Process for Interactive Systems, called UPi (Sousa & Furtado, 2004) and follows the approach to design UI prototypes dependent on the device (Coyette, Faulkner, Kolp, Limbourg, & Vanderdonckt, 2004).

Each phase is directly related to a specific area. This way, usability engineering supports the verification and validation in the inception and elaboration phases, software engineering supports verification in the construction phase, and semiotic engineering supports validation in the transition phase.

Figure 1 illustrates these four phases in the UPi-Test, each one with its flow of activities, artifacts, and techniques.

UPi-Test includes the verification and validation of usability and functionality of interactive systems UIs. Nonfunctional requirements are not in the scope of this work, such as aspects related to database (e.g., connection, integrity, etc.), security, and architecture. These aspects can be supported in future versions of this process.

Phase I: Inception

The inception phase is important in guaranteeing that the following phases achieve results to attend to users’ and customers’ usability goals. This phase has the constant participation of users in order to understand their requests, which are verified and validated according to usability engineering to allow the development of these requests. The description of the activities and artifacts in this phase is presented as follows.

Talk with Users and/or Customers

This activity consists of the first contact with users or customers, in which system analysts understand their profiles, objectives, and the scenario where they are included. In this activity, it is necessary to use a technique to elicit requirements.

We propose an initial informal talk in order to better understand users’ environment. Then, we suggest the performance of interviews, with the use of a questionnaire that aids in the identification of users and/or customers and their intended goals, preferences, and possible constraints.

Obtain Users’ and Customers’ Preferences and Constraints

In this activity, system analysts aim at eliciting users’ and customers’ preferences and constraints in order to design UIs that attend their needs and also to help usability experts during evaluation and to help UI designers during prototyping. Some examples of preferences are: design colors, font styles, navigation schemes. Some examples of constraints are: technology constraints, such as platform, device, and so on.

Consider HCI Models and Use Cases

Use cases represent an artifact from software engineering that identifies the functionality of interactive systems; users’ interactions, expressed through users’ tasks; and the system’s responses to perform these interactions (system’s tasks).

Task models detail a use case or a group of related use cases by specifying users’ tasks and system’s tasks. It is useful to support UI designers in the elaboration of prototypes because it is easy to identify the necessary views and objects in the prototype from the task model hierarchical structure.
Figure 1. Evaluation process of UIs
Analyze Existing Systems

In this moment, it is important to analyze existing systems. This involves a comparative study of similar systems. These systems can be used as a reference in order for system analysts to propose new functionality or to choose design patterns, which are all useful for the design of UI prototypes.

Consider Paper Sketches

After talking with users and customers, the UI designer designs paper sketches. This prototype is verified and validated, as presented in the next activity.

Verify with Expert

This activity consists of the verification of paper sketches by the usability expert. The expert is concerned with verifying if users’ goals, in terms of functionality, were included in the prototype, as well as if usability principles were used.

We propose that experts use the heuristic evaluation approach (Nielsen, 1993) as an inspection method. The usability heuristics will guide the usability expert in the process of verifying UIs’ quality of use. We also suggest the use of the task model to verify if all the specified functionality was designed in the prototype.

When the expert notices that a functional requirement or any usability principle is missing, change requests can be made, which leads to changes in the models and new proposals of prototypes.

Validate with Users and Customers

After the verification with the expert, we propose a validation with users and customers so they can approve the generated artifacts. If the prototypes do not attend users’ needs, change requests can be made, which leads to changes in the models (if new functionality is requested) and new proposals of prototypes (if changes in the navigation are requested). This process is repeated until the generated prototype attends users’ preferences and needs.

This activity early in the process provides flexibility for users and customers to evaluate the evolution of the system, therefore, designers and users feel more confident with the UI design.

After the conclusion of the inception phase, the resulting artifacts are verified and validated as paper sketches.

Phase II: Elaboration

The elaboration phase is concerned with designing and evaluating drawing prototypes. In this phase, we use verification and validation techniques, such as heuristic evaluation and validations with users. After this phase, the resulting artifacts are drawing prototypes validated according to usability requirements and patterns. The description of the activities and artifacts in this phase is presented as follows.

Use Design Patterns

In order to guarantee the quality of the product and efficiency of the project, we suggest the use of design patterns for graphical UIs. These patterns will guarantee that we elaborate and develop UIs following already verified and validated parameters, which can be incremented by the reports generated in the end of the transition phase.

Consider Drawing Prototypes

UI designers are responsible for designing drawing prototypes based on paper sketches previously validated, and on usability patterns. These prototypes are verified and validated by the following two activities.
Verify with Expert

This activity consists of verifying the usability of drawing prototypes by experts. We propose that experts use the heuristic evaluation approach (Nielsen, 1993) as an inspection method to verify whether or not certain usability principles are present in the prototype.

When the expert notices that any usability principle is missing, change requests can be made, which leads to new proposals of prototypes. At this moment, it is not necessary to make changes in the models because this approach evaluates the quality of use, not functionality aspects, which were evaluated in the previous phase.

Validate with Users and Customers

After verification with the expert, a validation is proposed to users and customers so they can approve the generated prototypes. At this moment, users and customers evaluate the used usability patterns and the style guide. This process is repeated until the generated prototype attends users’ preferences and constraints. If the prototypes do not attend users’ needs, change requests can be made, which leads to new proposals of prototypes.

After the conclusion of the elaboration phase, the resulting artifacts are verified and validated drawing prototypes, which support development, tests, and deployment activities.

Phase III: Construction

In this phase, the UI is developed and the application is integrated with it. Considering software engineering, we propose functionality tests of an executable prototype (i.e., a product with some functionality) or the final product (i.e., a product with all the functionality), using functional test cases.

The activities and artifacts in this phase are presented as follows.

Implement and Integrate

These activities consist of developing the UI and integrating it with the application. The integrated product, either a prototype or the final system, can be useful for evaluating the navigation, interactivity, and functionality aspects.

Elaborate Test Cases

Test cases can be elaborated starting in the inception phase, using paper sketches, they can then be updated in the elaboration phase, using drawing prototypes, and finished in this activity. This artifact focuses on the system functionality, not on nonfunctional aspects.

The technique used to define test cases includes the following topics: association to a use case, specification of the item to be tested, preconditions to execute before testing, identification of valid and invalid inputs, and the expected outputs. The actual outputs are compared with the expected outputs described in the test cases and this comparison is used as validation of the use case.

Verify with Expert

This activity consists of the verification of the functionality of the product by usability experts and developers. Examples of aspects that are verified are: consistency of the outputs, navigation, existence of error messages, results after clicking on objects, as well as other aspects identified in the test cases.

After this verification, developers and experts can generate change requests to correct the errors; which leads to the repetition of the implementation and integration activities.

After the conclusion of the construction phase, the resulting artifact is the integrated product, which is tested with consideration to usability aspects in the transition phase.
**Phase IV: Transition**

This phase comprehends, in general terms, the preparation of the test environment, which can be a test laboratory or the environment where the system is used. With consideration to semiotic engineering, we use some validation techniques and artifacts. The description of the activities and artifacts in this phase is presented as follows.

**Prepare the Environment**

To prepare the environment, we suggest the installation of the system, software for capturing the system’s use, and equipment, such as video cameras and necessary hardware devices. We also suggest the creation of questionnaires and checklists.

The test laboratory must be similar to the real user environment — with consideration to physical structure, climate, sound aspects, and equipment — in order to allow users to live the same conditions of the real environment. There should be a room where the test takes place and another one for observation.

**Install the Product**

In this activity, the product is installed, either a partial version or the final version of the system, in order to allow users to use the system in their real environment. This installation allows the tests to be performed.

**Usability Tests “with Users”**

This evaluation is performed with users. In it, evaluation techniques, proposed by semiotic engineering, are used, such as: recording, observation, questionnaires, and so on.

Before starting the tests, the usability expert talks with the user in order to: clarify that the system is under evaluation, not him/her; present the scenario used for the test; and make the user feel comfortable; which are aspects that influence the final results.

Observers that are in the observation room should fill out the questionnaires and checklists.

This activity can be divided in two moments, the first one, when the user interacts with the system to perform a task of his/her own interest; the second one, when the expert requires the user to perform a specific task.

In this activity, navigability, interactivity, and acceptability will be evaluated.

**Use Observations/Checklist/Questionnaires**

Observations, questionnaires, and checklists are artifacts and techniques proposed by the semiotic engineering in order to verify the user-system interactivity and communicability. Experts and observers will use these artifacts during the tests, which result in the definition of the quality of the interactive system. These results can lead to change requests for developers to correct the detected mistakes.

Users’ comments and the actual execution of the tests will be recorded to help in the analysis of the results of the questionnaires and of users’ observations.

**Make Corrections**

In this activity, developers make corrections proposed by experts after the tests. After the changes are made, users validate the product.

**Deliver the Evaluated Product**

As a result of the process, we have the results of evaluations, which are useful for future versions; and we also have a verified and evaluated product according to a set of techniques proposed by usability, software, and semiotic engineering.

If the product is a final version of the system,
it is ready to be delivered for use. If it is a partial version (e.g., executable prototype), the professionals need to perform the activities in the construction phase, then in the transition phase, until the product reaches its final version.

Elaborate Reports

The results obtained will be used as a basis for the elaboration of evaluation reports, which propose adaptations in the used patterns and in the creation of new patterns that can be used in future iterations.

CASE STUDY

In this chapter, we describe the case study of this research work, which is concerned with the evaluation of UIs for the SBTVD project, focusing on the applications: electronic portal, insertion of texts, and help.

Introduction

The digital TV represents digital and social inclusion for a great part of the Brazilian population, especially for people less privileged, who do not have access to computers, and therefore, cannot access the Internet.

The SBTVD must be adapted to the socioeconomic conditions of the country, as well as allow the use of conventional TV sets already in large use in the country in order to decrease risks and costs for the society.

The digital TV creates various possibilities of interaction between the TV and the user, such as: exchange of text or voice messages, virtual chats, searches for a favorite show, access to information about the government, and so on. These possibilities are different from the characteristics of the conventional TV, in which the user plays a passive role.

In the following section, we describe the performance of the activities proposed by UPi and UPi-Test. It is important to point out that the SBTVD is still under development. That is the reason why we cannot demonstrate all the activities of the process.

Phase I: Inception

Talk with Users and/or Customers

This activity was difficult to perform in this project because there is an almost unlimited number of users and/or customers. Fortunately, there were specific and clear specifications, established by the Brazilian government. Such specifications were used as requests from users and customers.

To define the scope of the application under our responsibility (access portal), we had meetings with representatives of the government and with other institutions that are participating in the project. These meetings were supported with brainstorming and the resulting decisions were analyzed and were shared with all the institutions through e-mails and discussion lists.

After these meetings, we decided that the portal will consist of a main application that allows the access to all other applications in the SBTVD, which can be: electronic mail, electronic commerce, EPG, help, electronic government, and so on.

Obtain Users’ and Customers’ Preferences and Constraints

The SBTVD project is concerned with various users’ profiles, including the ones who are and those who are not used to technology, but not including the ones with any kind of disabilities.

In order to identify their possible preferences and constraints, we studied existing systems; we had many meetings and workshops. The opinions of all the participants in the project were taken into consideration because we can also be considered to be potential users.
Our main goal in the usability workshops was to choose the best usability pattern for each requirement based on the evaluation of positive and negative aspects of each proposed usability pattern, as specified in the UPi activity define UI plan.

These proposed usability patterns were selected from a list, such as the one available in Welie (2005), which are organized in the following format: problem to be solved, solution, context in which it can be used, and graphical illustration.

For the participants to evaluate the usability patterns, we provided a set of guidelines. After they read the guidelines, each group evaluated the positive and negative aspects of each usability pattern suggested for the requirement. Then, the participants discussed and reached the final decision as to what was the best usability pattern for the requirement under discussion.

The personalization group considered the guidelines as they evaluated the positive and negative implications of each usability pattern, aiming at achieving Nielsen’s usability goals. For demonstration purposes, we focus on presenting the personalization of colors.

As a result of the evaluation, the personalization group decided to use color templates because of the greater impact of positive implications over the negative ones. The result of this work was the generation of a document associating usability patterns with their positive and negative implications. For instance, the selected pattern with more positive implications was “Provide predefined color templates to change font/background colors” instead of the pattern “Offer a list of colors from where the font/background colors can be chosen” because of the following positive implications: efficiency of use, compatibility of the system with the real world, and clarity of information. The second option had more negative implications than the first one, such as difficulty in use by beginners and the constant need to resize the space reserved to the presentation of information.

We defined that the applications in the SBTVD need to have a high level of usability. There was a list of technology constraints, especially the ones related to memory and processing capacity. Besides that, graphical representations need to be simple because conventional TV sets do not support images as computer monitors do, and the TV luminosity is very different from the one in monitors; consequently, colors and texts appear differently.

Consider HCI Models and Use Cases

In this activity, analysts and usability experts elaborated use case and task models, and then changed them, when change requests were made after the evaluation of users in the end of the Inception phase. The requests resulted in changes in the models because they were changes in the functionality, such as: do not consider access to multiple applications simultaneously (technical constraint) and include personalized help. These changes reflected in the models and will be considered in the prototypes.

Analyze Existing Systems

In this activity, we researched on the Internet in order to find digital TV systems already in use. It was very difficult to find them because there are few systems available for access in the Internet. This analysis was used to identify some UI design patterns for the digital TV, such as: upper and bottom options bar, menu on the left, navigation icons, interaction buttons, and so on.

Consider Paper Sketches

After the analysis of existing systems, UI designers elaborated paper sketches (Figure 2). However, they needed to be redesigned after the evaluation of users in the end of the inception phase, which resulted in change requests.
Some change requests resulted in the following changes in the prototypes: transfer the bar from the top to the bottom of the screen, include the “TV” button in the bar when accessing an application or in the portal; include the “Portal” button in the bar when accessing an application, take off activated icons (because of technical restrictions), and give a preview of the requested options of personalization. However, other requests resulted in changes in the models as well as in the prototypes, such as the ones mentioned in the activity “Consider HCI models and use cases.”

Verify with Expert

In this activity, we performed heuristic evaluations with three experts who considered the usability of the paper sketches. They observed aspects, such as: navigation between views and if usability principles were applied. Besides that, the experts verified if all users’ goals, in terms of functionality (in the task models), were included in the prototypes.

Validate with Users and Customers

In meetings with the institutions participants of the project and with representatives of the Government users evaluated the elaborated prototypes.

Various change requests, as well as users’ preferences were identified. These requests resulted in updates in the models and in redesign of the paper sketches.

As examples of preferences, we point out the following: Include the “TV” and “Portal” in the remote control; the bottom options bar should overpass the application; and make available a set of options for inserting text, such as the one used in computer keyboards, the one used in mobile phones, and in alphabetical order.

Phase II: Elaboration

Use Design Patterns

In this activity, the UI design patterns, identified while analyzing existing systems, were evaluated and the UI designers started to design drawing prototypes using the design patterns.

Consider Drawing Prototypes

The UI designers used image editors, such as Photoshop® and Corel Draw®, in order to design the drawing prototypes (Figure 3), which followed

Figure 3. Drawing prototype: Personalization
the selected patterns, the updated task models, and paper sketches. In this prototype, the small view on the right shows the preview of the required personalization about the color of the screen before applying it.

Verify with Expert

The drawing prototypes were evaluated by experts (Figure 4), who observed various aspects, such as: layout, colors and fonts, UI design patterns, and usability principles applied.

Validate with Users and Customers

This activity is yet to be fully performed in this project, currently; we have finished the document that specifies the SBTVD graphical UI, which is going to be evaluated by representatives of the Brazilian Government.

We scheduled a workshop with the other institutions participants of the project to evaluate the drawing prototypes (Figure 5). They requested us to elaborate other alternatives of UIs (including association of options in the UI with options on the remote control), increase the size of the options and fonts in the upper bar, and change the way to differentiate the selected option.

Phase III: Construction

Implement and Integrate

The implementation of the portal has not started yet. However, the developers have been studying and implementing simple applications for the digital TV, such as three possibilities of insertion of text (the one used in computer keyboards (‘qwerty’), the one used in mobile phones, and in alphabetical order).

Elaborate Test Cases

The elaboration of functional test cases has started since the inception phase, when use cases and paper sketches were elaborated. Functional requirements were selected and the associated test cases were elaborated: structure applications, access applications, access help and personalization (Table 1).

Verify with Expert

In this activity, the programmers and the experts verify the applications using the test cases for guidance. Some of these evaluations were done in the small applications developed for insertion of text.

Figure 4. Verification with expert

Figure 5. Validation with users
**Table 1. Test case: Personalization**

<table>
<thead>
<tr>
<th>Test case</th>
<th>Personalization</th>
</tr>
</thead>
</table>
| Test items | Changes in font and background color  
Changes in font size |
| Pre-conditions | The user must be accessing the portal |
| Inputs | The user must select a type of personalization: Font color The user selects green |
| Expected results | The portal must present the personalized content with a green background color |

**Phase IV: Transition**

The usability experts and the developers have prepared the test environment. For that, they have: prepared the physical structure (e.g., TV, video camera, couch, computer to simulate the set-top box, etc.) and installed the necessary software (i.e., software to capture the user interaction with the TV, Linux Operating System, Java Virtual Machine, and applications to be tested).

A group of three usability experts were responsible for performing the following activities before the tests started: First, they defined a questionnaire to apply with users in order to understand their characteristics and familiarity to the DTV technology. Second, they selected appropriate metrics (e.g., number of errors, number of access to the help, etc). Third, they created a checklist based on Nielsen’s usability goals and on the metrics from the previous step. Fourth, they prepared the environment with a DTV, a couch, and a center table in order to make users feel at home. Fifth, they selected ten users with different profiles between the ages of 18 and 26.

When the users arrived, each one at a time was taken to the DTV room, where a usability expert explained the goals of the test; applied the questionnaire; and defined a specific goal to be achieved while interacting with the DTV. While the user interacted with the application, usability experts filled out the checklist in the visualization room, where we monitored the environment and the user with a camera and captured the interaction with the DTV using specific software.

**FINDINGS AND FUTURE WORKS**

After the tests, we evaluated the checklists and generated reports with solutions to the problems encountered during the tests. These reports contain comments about the icons, the navigation, and the help module of the portal application.

- No user realized the possibility of navigation through the numbers associated to each icon, which represents an interactive application (see Figure 6).
- When they were told about this, they used it and said it was very practical.
- For them, the icons were not a natural representation, but they did not want to give any opinion about possible solutions.
- Some users did not notice the application that was selected. They said it was better to have a big square around the icon instead of the current selection.
- When the users were confused and wanted some help, they did not look for it in the portal (the blue option in the bottom bar that activates the help module). Some users activated the menu button of the remote control and others waited for the evaluator to tell them what was necessary to do to obtain the help.

![Figure 6. The portal application](image)
Integrating Usability, Semiotic, and Software Engineering into a Method for Evaluating User Interfaces

As the bottom bar is used as a design pattern to put the navigation options, and as each option is different depending on the situation (the application being executed), the users got very confused. We realized they looked at this bar only once, and then they did not look at it any more. After memorizing the options, they wanted to use them all the time, but the color was different, so they made many navigation errors.

From the execution of these tests, we were able to quantify the results: 85% of the users found the system easy to use, easy to read, and easy to navigate, but on the other hand, 50% had difficulties in identifying that an option was selected and in getting out of an unexpected situation.

Our next step is to make the necessary changes that are related to improving the layout, color contrast, and icon selection. In addition, the next tests will be done with the elderly in order to investigate people who are not used to interaction devices.

CONCLUSION

In this chapter, we present a new method, which focuses on integrating different techniques of usability, semiotic and software engineering. The aims are to design usable UIs following a model-based UI design process and to facilitate the test process by using the evaluated HCI models. In this manner, we hope to contribute to the development of interactive systems that are easy for users to learn and use, and to help testers in performing their usability tests in an efficient manner.

As main contributions, we focus on evaluating the usability of UIs with the constant participation of users and customers. Besides that, the integration of various approaches results in positive outcomes for the prototypes, as well as for multidisciplinary team members, who are better integrated and can have their knowledge enhanced, since they are continuously exchanging information and experiences.

REFERENCES


*This work was previously published in Verification, Validation and Testing in Software Engineering, edited by A. Dasso, pp. 55-81, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
ABSTRACT

How can we design technology that suits human cognitive needs? In this chapter, we review research on the effects of externalizing information on the interface versus requiring people to internalize it. We discuss the advantages and disadvantages of externalizing information. Further, we discuss some of our own research investigating how externalizing or not externalizing information in program interfaces influences problem-solving performance. In general, externalization provides information relevant to immediate task execution visibly or audibly in the interface. Thus, remembering certain task-related knowledge becomes unnecessary, which relieves working memory. Examples are visual feedback aids such as “graying out” nonapplicable menu items. On the contrary, when certain needed task-related information is not externalized on the interface, it needs to be internalized, stored in working memory and long-term memory. In many task situations, having the user acquire more knowledge of the structure of the task or its underlying rules is desirable. We examined the hypothesis that while externalization will yield better performance during initial learning, internalization will yield a better performance later. We furthermore expected internalization to result in better knowledge, and expected it to provoke less trial-and-error behavior. We conducted an experiment where we compared an interface with certain information externalized versus not externalizing it, and measured performance and knowledge. In a second session 8 months later, we investigated what was left of the participants’
Guidance in the Interface

knowledge and skills, and presented them with a transfer task. The results showed that requiring internalization can yield advantages over having all information immediately at hand. This shows that using cognitive findings to enhance the effectiveness of software (especially software with specific purposes) can make a valuable contribution to the field of human-computer interaction.

INTRODUCTION

Humans interact with information in the world around them by taking it in, processing it, and outputting reactions. To process information, they use cognitive skills such as thinking, learning, reasoning, recognizing, and recalling, as well as metacognitive skills, which entail thinking about cognition (for instance, planning, strategizing, or choosing between reasoning or calculation types). Cognitive science studies these domains of human thought. Much research in this field is done through the analysis of subject reactions to presented information. This makes cognitive science a source of knowledge that could— and does—guide interface and system designers toward a more effective presentation of information in computer systems. We believe that utilizing cognitive findings to enhance the effectiveness of software can make a valuable contribution. Increasingly humans exchange information with the aid of computers, for instance, in education, entertainment, office tasks, information search, e-mail, and many other domains. Advances in computer and multimedia technology ensure that the format of this information is increasingly diverse using multiple media. Moreover, programs can have hundreds of functions. However, progression becomes difficult with this complexity of choices and representations. Harnessing this complexity to make it manageable for humans gave rise to the domain of “usability.” Soon, among other things, the importance of minimizing user memory load became apparent. This resulted in recommendations to simplify the interface, restricting available options to those needed to carry out the task at hand, and to keep options visible on the interface so users could interact on the basis of recognition rather than recall (Nielsen, 1994). In other words, the aim was just-in-time delivery of just the right information, obviating the need for memorization and extensive search in memory.

Our research does not aim at uncovering more principles that make systems even more usable, intuitive, or appealing. It goes beyond plain usability and focuses on how to shape interfaces that induce a user to learn cognitive and metacognitive skills, and thereby learn about the domain underlying the interface. We would like to find patterns of human behavior occurring with computer use, to find out what kind of behavior certain decisions in interface design provoke, not only during interaction, but also after delays and in transfer situations. In this, we feel that one continually has to consider the real purpose of the system. If a system ought to teach material to students or children, or needs to make sure that users do not mindlessly follow interface cues because the task to perform is of a certain crucial nature, then we should know what it is about an interface design that induces people to think and learn. In this chapter, the focus is on internalization and externalization of information, and how this may lead to different behavior on the user’s part. In the following sections, we explain the different terms used in this context. After this, we will look at the pros and cons of externalizing and internalizing information, and some effects of varying interface elements on learning and metacognitive processes. In the next sections we discuss an experiment on users’ behavior that two interface styles (internalization and externalization) provoke, and more specifically, the amount of planning and learning from the users’ side. In the concluding section, we discuss our findings and lay out our future plans.
EXTERNALIZATION VS. INTERNALIZATION INFORMATION: RELATED FIELDS

Visualization of task-specific information, thus minimizing user memory load, is often called “externalizing” the information (Zhang & Norman, 1994). Externalization of information can be contrasted with internalization of information, whereby certain task-related information is less directly available and needs to be internalized (inferred and stored in memory).

Early work of Simon (1975) examining advanced chess skills and strategies to solve the Tower of Hanoi puzzle had noticed the interdependence of “external memory” (perception, cueing recognition memory), and internal memory (recall memory). Norman (1988) argued for the need of a similar analysis for human-computer interaction. Tabachneck-Schijf, Leonardo, and Simon (1997) created a model in which distributed cognition was reified. For an example of distributed cognition, showing the interaction between internal and external memory in creating a Microsoft PowerPoint page, see Tabachneck-Schijf (2004). The question is, how much information should be internalized and how much should be externalized for an optimal task execution? Some research shows that the more information is externalized, the easier executing a task becomes (e.g., Zhang & Norman, 1994). Other research shows that externalizing all needed information seduces the user into mindlessly following the interface cues and learning or planning little (e.g., Mayes, Draper, McGregor, & Oatley, 1988). Yet other research shows that giving users incentives to plan and learn (i.e., internalizing information) pays off in better performance (e.g., O’Hara & Payne, 1999). Internalizing and externalizing information are related to, but not the same as, internal and external memory, internal and external cognition, and plan-based and display-based problem solving. All relate to having information in the head, be it in long-term or in working memory, versus information that is available via (direct) perception. Simon and colleagues refer to external memory as all information that can be made available for perception (e.g., Greeno & Simon, 1974). In practice, this comes down to all information that, within the task execution, can be made available for direct perception and working memory. Internal memory is all the information within long-term memory that can be made available to working memory.

External cognition is a term coined by Scaife and Rogers (1996) and refers to the interaction between internal and external representations (the latter being “external memory”) when performing cognitive tasks (e.g., learning). Scaife and Rogers provide a useful overview on the use of graphical representations in educational environments and outline a theoretical framework. They conceptualize how different kinds of graphical representations (e.g., diagrams, animation, and multimedia) are used during cognitive activities. The framework presents a set of core properties, of which the central one is computational off-loading (the extent to which different external representations reduce the amount of cognitive effort required for cognitive activities). This property was also mentioned by Greeno and Simon (1974) and Larkin (1989).

Plan-based and display-based problem solving relate to which source is used to draw information from. In plan-based problem solving, one uses detailed problem-solving strategies from (internal) long-term memory. Display-based makes little use of learned knowledge but relies on (external) interface information (Larkin, 1989; O’Hara & Payne, 1999).

Externalizing and internalizing information, finally, refer to making information available on external memory versus not making it available on external memory, thus requiring recording it on internal memory. We feel that studying the relationship between internalizing and externalizing information in order to find effects of interface
information on behavior and problem solving is an important first step.

**PROS AND CONS OF EXTERNALIZING INFORMATION**

In many instances, and for many purposes, externalizing information and thereby reducing working memory load is indeed a fine idea. Externalizing interfaces “take the user by the hand,” limit choices, and provide feedback. What information should a computer system offer? If one is working with a dynamic visual display on a problem-solving task, problem states are directly visible and elements that are normally stored in working memory can now be read from the display. A way to facilitate reduction of memory load is by visually showing what possible steps can be taken to go from one state to the other in the problem space (for a problem space, see Figure 4). It prevents users from internalizing a great deal of information. Other ways to reduce memory load and support explorative behavior are, for instance, guidelines such as “keep the number of possible operations small,” “make the possible operations distinguishable,” “make clear what the consequences of every action will be,” “make the effects of actions visible once they have been executed,” or “make actions easily undoable to make it safe to experiment” (Lewis & Polson, 1990; Van Oostendorp & De Mul, 1999).

Systems such as cash dispensers, automated airline check-in systems, and so forth should be designed in such a way that users can simply and easily follow the interface cues. Other examples are wizards and help options. Graying-out menu items that cannot be used at that moment is also an example of externalizing information. For example, in Word, you cannot select “paste” from the “edit” tab in the menu when nothing is copied or cut first. “Paste” is shown in gray, indicating that the command exists, but that using it is impossible at the moment. In various lists of user interface guidelines similar issues, such as “visibility status” and “visual feedback” are mentioned.

Interfaces that externalize do minimize working memory load. For instance, Larkin (1989) considered the role played by differences in external displays in cognitive problem solving, finding that making information visible, enabling “display-based problem solving,” helped people recover from interruptions in work. As mentioned,

---

**Figure 1. Examples of externalization by graying-out items**

![Example of externalization by graying-out items](image-url)
Guidance in the Interface

Zhang and Norman (1994) showed that the more information was externalized, the easier solving a problem became.

Van Joolingen (1999) studied the idea of learners actively engaging in order to learn about the properties of what is underlying a certain domain. He mentioned the reduction of memory load in the context of research on the role of learner support in discovery learning. Discovery learning assumes that learners take an active role and construct their own knowledge base by interacting with a domain, and inferring rules from the results of these interactions. Active involvement of the learner would result in a better base of knowledge in the learner as opposed to more traditional ways of learning. The question is, of course, how to accomplish this behavior from the users’ side. He addressed the issue of supporting discovery learning by means of cognitive tools. Cognitive tools were defined as instruments that support cognitive processes by relieving working memory and which can aid the learner to direct the process, to perform a part of it, or to externalize part of processes and memory.

However, externalization of information also appears to have disadvantages. There is some evidence that during computer-based problem solving, with most information externalized users simply “follow the yellow brick road.” Users search the interface for cues and reliably find them, perform an action that alters the interface, thus enabling another search for cues and so, applying simple perception-based means-ends analysis, eventually complete the task. If the task appears to be completed correctly, then it follows that users will not be triggered to plan, learn, or look for underlying rules or reasons, and also, that people can only carry out a sequence of actions if the interface supplies the needed cues. In a study comparing memory of users for a sequence of actions, researchers found just that (Mayes et al., 1988). Users using a command-style interface (e.g., LaTeX) with most cues externalized could not. Also, external cues can easily lead subjects to incorrect actions. Tabachneck (1992) found that when the needed information was externalized on a graph so that simple perception yielded the correct answer, subjects gave the correct answer. However, in different representational formats simple perception yielded an incorrect answer; here subjects also followed the interface cue and gave the incorrect answer. In summary, externalizing information does decrease memory load, but drawbacks are that people are likely to perform continual trial-and-error recognition-based search and not learn or plan. Also, the external information may lead to the wrong answer and thus to errors. Task execution via externalized information may also be very resistant to speeding up. We feel that there are many tasks for which, during task execution, internalization of information is desirable. Examples are tasks with the purpose to acquire knowledge or skills, be they cognitive or metacognitive; tasks one does often, and consequently speeding up and reducing errors is desirable (such as booking a trip for a travel agent), or tasks where the cost of errors is high (e.g., controlling nuclear power plant or flying an airplane). When looking more specifically at problem-solving tasks, planning and learning are thought to be essential factors. It is therefore desirable to do more research on how interface design, internalizing, and externalizing information can stimulate people to do more internalization. Not externalizing all information may stimulate self-directed search, more active planning, and learning.

VARYING ELEMENTS AND (META)COGNITIVE PROCESSES

O’Hara and Payne (1998) and Trudel and Payne (1996) found that when people are induced to plan, too strong a reliance on external information
leads to negative effects regarding planning and transfer of skills. Inducing people to do plan-based problem solving would lead, they hypothesized, to internalization of metacognitive and cognitive knowledge, in turn improving task performance and easier recovery from errors (see also Payne, Howes, & Reader, 2001). O’Hara and Payne’s (1998) experiments showed that plan-based activity did lead to shorter solution routes and easier recovery from errors, while a display-based strategy involved more steps because of more (trial-and-error) searching. O’Hara and Payne used an interesting methodology to get people to plan more. In one of two conditions, they made the interface harder to use by imposing delays on operators. They reasoned that subjects plan more when there is something at stake: here, reducing costly delays. A similar observation was made by Svendsen (1991). Using the Towers of Hanoi problem, a high-cost interface yielded improved understanding of problems. In the studies of Payne and colleagues, the amount of information externalized was not varied. We felt that not externalizing all information will similarly stimulate students to self-initiate a search for the missing information to plan and to learn.

In the first session of our experiment, discussed next, we researched whether varying the amount of externalized information would have an effect on performance, planning, and learning while solving a problem. In the second session we assessed what was left after 8 months of what subjects had learned in the first session and whether this still had an effect on task performance. Details follow.

**EXPERIMENT SESSION 1**

**Hypotheses**

1. **Externalization will initially yield better task performance than internalization.** Knowledge is yet to be acquired, so guidance by indicating the legality of moves will help externalization subjects.

2. **Internalization yields better task performance later, especially after a severe distraction.** After a while, subjects will have learned to solve the puzzle. Internalization subjects had no guidance and had to acquire the solving skill by themselves. They will have stored the rules in long-term memory more solidly, and have the needed information more readily available and thus perform better later, especially after an interruption erased working memory. Because of the guiding nature of the interface, externalization subjects will plan and think less than the internalization subjects, therefore work more on the basis of trial and error and consequently display worse performance.

3. **Internalization yields better knowledge.** Not having externalized information available will motivate a subject to start planning on the basis of self-acquired rules. After the
experiment, we expect the explicit knowledge of rules to be memorized better by internalization subjects. There may also be a difference in implicit procedural knowledge, but we expect it to be smaller since both groups will have played the puzzle a similar number of times.

An experiment was conducted in which subjects solved a puzzle on a PC nine times in two conditions: internalized and externalized. In terms of performance, we expect the outcomes to take the following course:

- **Phase 1:** Three trials. Subjects do not have any needed knowledge available yet, externalization offers more interface cues, and will result in better performance.
- **Phase 2:** Three trials. Performance in the two conditions will be equal by this time, because internalization subjects acquired internal knowledge that compensates for the interface cues.
- **Interuption:** This task is sufficiently demanding that all of the working memory is erased.
- **Phase 3:** Three trials, after the interrupting distraction task. Internalization subjects perform better because of better internalized knowledge and a more elaborate plan.

**Materials**

Our problem-solving task, “Balls & Boxes” (B&B), is an isomorph of the classic puzzle “Missionaries and Cannibals” (M&C). Five missionaries and five cannibals are standing on a riverbank, and all have to reach the other bank by boat. Constraints are that the boat only holds three creatures, and the minimum to sail is one, because the boat cannot move by itself. Furthermore, the cannibals can never outnumber the missionaries at any place (except when there are zero missionaries), or the missionaries will be eaten. Our B&B problem (Figure 3) uses exactly the same problem space (Figure 4), but is more abstract.

We made the puzzle more abstract to reduce the many common knowledge elements in the classic M&C: cannibals eat people, boats cross rivers, boats cannot sail empty, and so forth. Using boxes, blue and yellow balls, and a dish instead, we offer a semantically less rich problem, avoiding the rules to be learned and remembered too easily.

The rules of B&B were as follows:

1. Balls should be transported using the dish.
2. You can transport at most three balls at a time.
3. To move, the dish must contain at least one ball.
4. The dish has to reach the boxes in turn.
5. There can never be more blue than yellow balls in the dish (except when there are zero yellow balls).
6. There can never be more blue than yellow balls left in either of the boxes (except when there are zero yellow balls).

Rules 5 and 6 are the most problematic rules, making no semantic sense whatsoever. In our puzzle, it was possible to consult all the rules, but having access to these does not imply that one knows how to solve the problem, and surely

**Figure 3. The Balls & Boxes puzzle**
Guidance in the Interface

not how to solve it in the most efficient manner. Below we depict the formal problem space of the puzzle (Figure 4). The numbers indicate the number of blue and yellow balls in the left and the right box, and the “∗∗” indicates on which side the dish is (see the legend of Figure 4). The shortest path from “A” to “N” takes 11 moves. There are also states of the puzzles that we call “dead-end states,” such as states D, E, G, J, K, and M. These states are not on a direct solution path, and force backtracking.

The interface controls of the puzzle were straightforward (Figures 5 and 6). To get balls into or out of the dish, blue or yellow up-arrows or down-arrows had to be clicked. To move the pink dish horizontally, one had to click a pink arrow (left or right). After the dish was moved, the balls automatically dropped into the box. To consult the rules one clicked on the rules tab. The rules screen overlaid the puzzle screen. The puzzle was designed in two versions:

1. **Internalized:** All arrows were always colored (Figure 5) and clickable, providing no information about the legality of moves (performing illegal moves is possible). One could click all buttons at all times; however, illegal moves would be carried out only partially. For instance, if one wanted to transport the dish empty from left to right and clicked the right pink arrow, the empty dish would move to the right. Then, an error

---

**Figure 4. Problem space of the Balls & Boxes problem**

![Figure 4. Problem space of the Balls & Boxes problem](image-url)
Guidance in the Interface

notification would pop up saying, “This is not possible.” By clicking “OK” in the dialogue box, the program would undo the move.

2. Externalized: The arrows are colored (and thus clickable) only when an action is legal. The arrows are grayed-out (unclickable, as in Figure 6) when a move is illegal. For instance, in Figure 6, moving the dish empty from left to right is illegal (violating rule 3), externalized by graying-out both pink arrows. It is now only legal to move balls up into the dish.

Subjects and Design

Thirty subjects of age 19–28 (mean age 21.8), experienced with PCs, were recruited at Utrecht University. Subjects were paid 5. The puzzle, a Java-applet, ran on a Pentium IV 2.0 GHz PC with a 17” monitor. Our design has one random-as-

Figure 5. Internalized version: No information whether an action is possible (no gray-out)

Figure 6. Externalized version: Information whether an action is possible (grayed-out)
assignment between-subjects independent variable: internalization versus externalization. Among the dependent variables were the following:

- Performance measures (logged)
  - Number of puzzles subjects solved per phase (maximum three per phase)
  - Speed: time needed to solve the puzzles
  - Reaching dead-end states. These are puzzle states that are off a direct solution path, indicating trial-and-error search (states D, E, G, J, K, and M in Figure 4)

- Knowledge test: After the puzzles, we measured implicit and explicit knowledge of the rules and shortest-path solutions of the problem, using a paper questionnaire.

- Attitudes: Likert-scale questions concerning perceived amount of planning, feeling lost during interaction, and other such measures.

**Procedure**

The experiment was conducted in the usability lab at the Center for Content and Knowledge Engineering, Utrecht University. We informed the subjects of the course of the experiment, and gave a brief oral explanation of the interface and a short demonstration. The experiment consisted of nine puzzle trials, divided into three equal phases, and a 10-minute distraction task between phase 2 and 3. The maximum time for each trial was set at 7 minutes. Slightly different starting situations of the puzzle were used to avoid subjects simply repeating actions (states A, B, and C in Figure 4). Also, in the second phase, the playing direction of the puzzle was reversed to right to left. In the third phase, the playing direction was set to left to right again. After the last trial, subjects filled out a knowledge test (score 0–8) consisting of four multiple choice and four open questions with screenshots of puzzle situations. They had to judge and explain whether and why certain actions were possible (implicit knowledge), and recall the rules (explicit knowledge). Subjects also rated how clear the rules were for solving the problem.

**RESULTS EXPERIMENT, SESSION 1**

Number of Puzzles Solved per Phase

Subjects could attempt to solve three puzzles per phase. A MANOVA (analysis of variance) showed a significant main effect of phase on the average number of puzzles solved, $F(2,56) = 53.74$; $p < .001$ (Figure 7) (pointer to figure). The number of puzzles solved (out of three) improved in later phases, indicating a learning effect. Post-hoc comparisons showed that only the difference between performance in phase 1 and phase 2 was significant ($p < .05$) (Figure 8) (pointer to figure). Although the graph suggests an overall better performance for internalization ($M = 6.93$, $SD = 2.05$ vs. $M = 6.33$, $SD = 3.09$), this did not reach significance. There were no interaction effects.
Guidance in the Interface

**Time Needed for the Puzzle Trials**

A MANOVA showed a main effect also for phase on average solving time, $F(2,42) = 35.16; p < .001$. The time subjects needed to solve puzzles lessened in later phases, also a learning effect. Post-hoc comparisons showed that all subjects needed significantly more time in phase 1 than in phase 2, and also in phase 2 than phase 3.

The graph suggests that internalization subjects need some more time on average, but ANOVA showed that this was just a trend ($F(1,21) = 2.79, p = 0.11$). Looking closer, it was only in the beginning, in phase 1, that internalization subjects needed more time ($M = 238.77, SD = 62.20$) to solve the puzzles than externalization subjects ($M = 186.77, SD = 67.08$). This difference was significant, $t(21) = 1.94, p = 0.03$ (one-sided). After this first phase, the differences were minuscule and no longer significant.

**Dead-End States**

Below is the number of times per phase subjects reached dead-end states.

A MANOVA showed an interesting nearly significant main effect of interface style (Figure 9). Overall, externalization subjects reached more dead-end states ($F(1,28) = 3.58; p = 0.06$). In addition, there is a trend for an interaction effect of phase and version ($F(2,56) = 2.11; p = 0.13$). Internalization subjects significantly improved from phase 1 to 2 ($M = 11.4, SD = 4.70$ and $M = 3.4, SD = 3.18$), $t(14)=5.80, p = 0.00$, one sided). This is an improvement of eight states. They also improved from phase 2 to 3 by almost two states, nearly reaching floor level ($M = 3.4, SD = 3.18$ and $M = 1.47, SD = 2.07$), $t(14)=1.96, p=0.04$, one sided. Externalization subjects improved by 6.2 states from phase 1 to 2 ($M = 12.67, SD = 6.91$ and $M = 6.47, SD = 7.11$), $t(14)=2.74, p = 0.02$, one sided. But after the interruption, it was different. From phase 2 to 3, they did not further improve. On the contrary, in phase 3 they reached more dead-end states than before, although not significantly so. Here in phase 3 the difference between internalization ($M = 1.47, SD = 0.53$) and externalization ($M = 7.73, SD = 10.64$) was significant, $t(28)=2.24, p = 0.04$. Externalization subjects reached more than six more dead-end states in phase 3 than internalization subjects.

**Knowledge Test**

Before measuring knowledge of the rules and states, we confirmed that there was no difference in the number of times subjects in both conditions consulted the rules. The knowledge test that
subjects received after the experiment consisted of several procedural knowledge questions and one explicit knowledge question. The Balls & Boxes puzzle contained more than 30 legal states (Figure 4). There were seven procedural implicit knowledge questions about seven of those states (open and multiple choice) in which subjects were visually presented with a puzzle state. They had to judge whether certain moves led to the solution, and explain why. The scores on those seven questions were high for both internalization (M = 6.3, SD = 0.96) and externalization subjects (M = 5.8, SD = 1.20), considering that the maximum score was seven. There was a trend that internalization on average scored higher than externalization on the procedural knowledge questions (t(28) = 1.17, p = 0.12).

To test the explicit rule knowledge, the possibilities were limited. The puzzle had only few rules, some of which (rules 1 to 4) were extremely easy to grasp and remember. Rules 5 and 6 are more crucial and define the difficulty of the puzzle. They can be merged to just one rule: “Blue balls can never be in the majority at any place, except when there are only blue balls (zero yellow balls).” Subjects were asked about this rule with a multiple-choice question. All internalization subjects answered this question correctly, whereas of the externalization subjects only 60% answered it correctly. This difference was significant (Phi = -.50, p < 0.01). Furthermore, subjects had to estimate the clarity of the rules. They rated the question “the given rules were clear enough to solve the problem” (score range 1–5). Internalization subjects found the rules clearer than the subjects from the externalization condition (M = 4.13, SD = 0.52 and M = 3.53, SD = 1.25), t(28)=1.72, p = .04 (one-sided).

**SUMMARY EXPERIMENT SESSION 1**

Our first hypothesis stating that initially externalization yields better performance was only partly supported in terms of the time subjects needed. Internalization subjects did take more time in the first phase. Other measures were in the expected direction, but not significantly so. Unexpectedly, externalization yielded no better performance in terms of puzzles solved in the beginning, as we expected.

The second hypothesis stating that internalization yields better performance in a later stage was partly supported. We saw that the time subjects needed stabilized, and the same was true for the number of trials solved per phase. However, we also looked at performance in a more delicate manner, not in terms of time or correctness, but at how, via what route, subjects reached their goal. We introduced the measure “dead-end states” to inform us as to how subjects behaved, in terms of the insight they had, the “smartness” of their route. We assumed that internalization subjects do some smarter, more elaborate planning, while externalization subjects are expected to solve more by trial and error and on the basis of interface cues. It showed that internalization subjects performed better after the interruption and reached those dead-end problem states less often in all three phases (indicating less “lostness”). Furthermore, there was also the trend-like interaction that after the interruption, internalization subjects kept improving, while externalization subjects fell back, reaching more dead-end states than they did before. This confirms our expectation that after the interruption, internalization subjects would continue to work on the basis of the plan-based strategy as they did before. Externalization subjects, on the other hand, perform no better after the interruption. They fell back to depending on the interface, having a less elaborate plan.

The third hypothesis in which we expected that internalization would result in having better knowledge was supported. We also assumed that internalization subjects, who could rely less on interface information, had to build a stronger, more elaborate plan. Though when testing implicit knowledge both groups scored equally
high, with a trend advantage for internalization subjects, when asked about explicit knowledge of the crucial rule that defines the difficulty of the puzzle there was a significant difference. All the internalization subjects could correctly answer this question, whereas only 60% of the externalization subjects could—in spite of having the rules readily available for consultation and consulting them as often as internalization subjects. Furthermore, there was also the tendency that internalization subjects rated the clarity of the rules higher. This is intriguing, because in the externalization version of the puzzle subjects had interface feedback and were able to consult the rules. Internalization subjects, who only had the rules and no interface help found the rules clearer. We carefully interpret the latter two findings as indicators of better understanding in the internalized condition.

**EXPERIMENT SESSION 2**

We were curious to see how stable this better knowledge provoked by internalization was, and therefore decided to conduct a second session after a delay of 8 months. We invited the same subjects of experiment 1; 14 of the 30 were available. There were two reasons for this rerun. First, to see whether the better knowledge measured in the internalization subjects had endured, we asked subjects to solve B&B five times (experiment 1 showed three to four puzzles suffice for all subjects to be able to solve the puzzle within the allotted time). Second, to see whether the better knowledge might result in better performance on a transfer task, we also confronted subjects with a transfer problem. Transfer problems require subjects to apply acquired skill on a different task of the same nature. To be able to measure differences in performance between the two initial groups (internalization and externalization) on the same material, we presented all subjects with the same material this time, one interface style, namely, externalization. Note that the internalization subjects had to make a change in interface style.

**Hypotheses**

1. Because internalization subjects still have better memory of the rules and solution, they will perform better. We expect the internalization subjects to solve the puzzle faster initially because of their better knowledge of the rules. They will be faster in recollecting knowledge needed to solve the puzzle. After a while, we expect the two groups to perform more or less equally on this puzzle, like in session 1.

2. Internalization subjects will perform better on the transfer tasks. After five times of B&B, we expect the two groups to perform at the same level. But when confronted with transfer tasks (that have similarities but also a few differences), we expect internalization subjects to perform better, again because they possess better knowledge of the rules.

**Materials**

To test knowledge retention, all subjects first solved the B&B puzzle in the externalized version five times. To test transfer performance, we used another puzzle of the same M&C “family,” but with varying characteristics. We first used a quite literal version of M&C, which was further away in terms of transfer.

**Balls & Boxes**

The only difference between this and the previous B&B was that the rules were not consultable anymore, so we could obtain a more pure measure of retention. All subjects received the same externalized version. This meant that attempting illegal moves (moves that violate rules) was not
possible, as the externalized interface only allowed legal moves.

Missionaries and Cannibals

The second problem was a transfer task, ironically the original version of the M&C. It was a game that literally shows missionaries, cannibals, a river, and a boat. The solution algorithm to this problem was the same as B&B, but there were some difficulties that did not exist before. The most important one is that there were not five of each entity as before, but three, and the playing direction was always from right to left (Figure 10). Although it might seem that one can solve the three-creature problem faster, this is not so—the shortest path to the solution is also 11 moves. Also, unlike in B&B, in this game attempting illegal moves does not lead to an error message, but subjects would “die,” as in a computer game, and the game would start over. Errors thus had more severe consequences, and subjects needed to exercise caution. Subjects had to solve the problem as many times as they could in 8 minutes.

Subjects and Procedure

Fourteen of the 30 subjects from session 1 were willing to participate a second time. Luckily, seven of them had worked with internalization, and seven with externalization in session 1. We confirmed that these 14 subjects were subjects who performed “normally” (no extreme scores) in session 1. The experiment consisted of two parts:

1. B&B (five times). The maximum time for each trial was set at 7 minutes. Slightly different starting situations of the puzzle were used to avoid subjects simply repeating actions.

2. M&C (8 minutes).

After completing the experiment, subjects received a 5 reward.

RESULTS EXPERIMENT SESSION 2

• Balls & Boxes. Encouraging results were found concerning solving the puzzle correctly again the first time. After not having seen the puzzle for 8 months, it took the internalization subjects only half the time used by the externalization subjects to solve the first B&B puzzle (M = 432 sec, SD = 314 and M = 778 sec, SD = 397). This difference was significant, t(12) = –1.81, p = .05. There were no further significant differences between the two groups. After the first puzzle, as expected, all subjects solved the remaining four trials of B&B puzzle equally well, just as 8 months ago.

• Missionaries & Cannibals. This puzzle was a bit further in transfer. The graphical characteristics of this puzzle differed considerably. Still the algorithm to solve it in itself was the same, but the number of creatures to transport was different, and also the maximum number of creatures allowed in the boat. Although the shortest solution
path is the same length as B&B, the problem does have a smaller problem space. The same basic concept had to be applied to a situation that differed at face value, and also underneath. Out of our 14 subjects, 10 managed to solve this puzzle one or more times. Just as in the B&B puzzle, internalization subjects solved it the first time faster (M = 176 seconds, SD = 72.5 vs. M = 302.8, SD = 202.7), though it was just a trend, t(8) = –1.44, p < 0.10 (one-sided). Moreover, internalization subjects managed to solve the puzzle three times more often (M = 4.83, SD = 1.94) in the 8 minutes than externalization subjects (M = 2.5, SD = 1.73). This was significant, t(8)= 1.94, p < .05 (one-sided).

SUMMARY EXPERIMENT SESSION 2

Eight months is a long time, but still the interface style subjects received in session 1 appeared to make a difference. Subjects were not told at the first session that they would be asked to play the puzzle again, and when we approached them, we explicitly told them it was not the same experiment they had participated in. Upon being confronted with the puzzle again, the subjects had to recollect the rules and solution strategy from their long-term memory.

The two groups both worked with an externalized version this time and had exactly the same information. We looked at how long it would take subjects to remember how to do this correctly. It showed that in solving the first B&B again, internalization subjects were indeed significantly faster, almost twice as fast, in fact. This is all the more remarkable given that these subjects received, to them, an unfamiliar interface—externalized rather than internalized. After that first success, the performance of both groups equalized. This is coherent with the results from the first session, where stabilization in performance also occurred. The above finding supports our first hypothesis of experiment 2 stating that internalization subjects would still have a better memory of the rules and solutions of the puzzle.

The M&C puzzle that subjects received was the same for everyone. It can be regarded as an internalization version (no direct interface clues were available, all controls were available all the time). Also here there were some interesting results. The subjects that worked with the internalization version of B&B managed to solve this puzzle three times more often than externalization subjects, and this was significant. Furthermore, just like in B&B, internalization subjects needed less time to solve the problem the first time, although this was just a trend. It supports the second hypothesis stating that internalization subjects will perform better on a transfer task. In summary, we demonstrated that the interface style subjects worked with 8 months ago still appeared to be of influence months later, both in solving the same puzzle again as in solving a far transfer puzzle. We take it as encouraging support for the better memory of knowledge provoked by the interface style subjects worked with.

GENERAL DISCUSSION ON THE TWO SESSIONS

We analyzed the influences that externalizing certain information on the interface, thereby making it visible, or not externalizing it has on performance and knowledge acquisition in a problem-solving task. Furthermore, we looked at the long-term effects after 8 months, and at the issue of transfer. Both in session 1 and in session 2, the delayed session, NOT externalizing information led to enduring advantages in performance and knowledge, while externalizing information merely led to a fleeting advantage at the very start of session 1.

According to Zhang (1997), the more information is externalized, the easier performing a task becomes. We hypothesize, however, that being
Guidance in the Interface

led by the interface also incurs a cost, namely, a lessening of metacognitive activity such as planning, and consequently, less learning. In such situations as where learning is required, transfer is at stake, or the task is prone to interruptions, planning, learning, and in general, getting to know more about the structure of the task at hand are desirable outcomes. Surprisingly, performance measures such as time needed to solve the puzzles and number of correctly solved puzzles were not much influenced by the interface style—if anything, the internalization group was at an advantage. Zhang's (1997) prediction was not confirmed. We feel that the attention of users in the externalization group was immediately taken by the guiding nature of the interface, and as they were not confronted with actual mistakes (one could not make illegal moves, only inefficient legal moves were allowed), they simply kept on solving without feeling a need to figure the problem out more. Subsequently, less active learning took place. This idea of attention taken by an interface fits with Carroll and Rosson's (1987) paradox of the active user—users of computer systems are so consumed with immediate productivity that they are not motivated to take time to learn better ways of accomplishing a task.

Contrary to the externalization group, the internalization group was confronted with errors for which an explanation was not immediately available. These subjects incurred a cost when making errors: a dialog box popped up which they had to click away, and their attempted move was reversed. Though both groups consulted the rules sheets equally often, this cost probably contributed to motivating the internalization subjects to study the rules better in order to avoid incurring the cost. We found that the internalization group possessed better explicit knowledge of the rules and engaged in more planful problem solving. Applying more metacognition to avoid a cost concurs with the findings of O'Hara and Payne (1998, 1999).

During session 2 we found that the effects found in session 1 were still present. Again, only the internalization subjects from months ago showed advantages. This group was both faster in solving the same problem again, and also faster in figuring out the solution to a transfer task. It is remarkable that the influence of working with one interface or the other has effects even after such a long time.

In general, it was surprising for us to see that externalization showed even less advantage than we thought, almost none, to be precise. We expected externalization at least to be of help in the beginning when users were not familiar with the system and the problem. It was, but only in time taken on the first few trials—and this was just a trend that was not confirmed in the number of puzzles solved correctly. This very small and fleeting advantage did not concur with Zhang's (1997) findings. Please note that our type of externalization parallels Zhang's more physical type: both restrict the number of illegal moves that can be made, but do not alter the problem space or give cues as to what the shortest path is through the problem space.

We find the results so far encouraging in that lessening the amount of externalized knowledge apparently can encourage cognitive and metacognitive behavior. Moreover, one can be very precise about which information not to externalize—in other words, this way the information to be learned can be manipulated. Therefore, we feel that the issues of manipulating the amount of externalization in interfaces deserve more attention. In the context of current technology and the widespread use of computer systems in virtually all domains, we are convinced that understanding how the system’s users will perceive, reason, and interact with the system can be crucial. The implications of the findings so far and the ones in the future can be valuable for development of applications where active engagement and learning from users is the aim. Examples are systems to teach material to students or children,
or situations where it is important that users are not “kept” stupid by the software. In certain situations, users of a system need to understand the underlying structure/rules of the system because the tasks are of a crucial nature. To accomplish this we have to look beyond plain usability and common sense and better scrutinize the interaction between internal memory and cognition on the one hand and external memory and recognition and perception on the other hand. We believe this can, for certain types of systems, enhance the effectiveness in achieving their targets.

FUTURE TRENDS

There are still many challenges in human-computer interaction and many issues that need to be explored. Understanding how users will react to interface information (on the basis of cognitive findings) is one important issue in attuning software to its purpose, thereby allowing it to achieve its goal. In the future, we will further investigate this issue by exploring behavior in different types of more realistic planning-related tasks. As a more realistic planning task, we think of, for example, spreadsheet or drawing applications where actions are less repetitive, more complex, and could be part of a real job. We are currently designing such an environment.

REFERENCES


*This work was previously published in Cognitively Informed Systems: Utilizing Practical Approaches to Enrich Information Presentation and Transfer, edited by E. M. Alkhafia, pp. 74-101, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 2.8

Ergonomic User Interface Design in Computerized Medical Equipment

D. John Doyle
Cleveland Clinic Foundation, USA

ABSTRACT

Current statistics suggest that preventable medical error is a common cause of patient morbidity and mortality, being responsible for between 44,000 and 98,000 deaths annually, and resulting in injuries that cost between $17 billion and $29 billion annually. An important approach to tackling this problem is to apply system design principles from human factors engineering (ergonomics). By doing so, systems and equipment become easier for people to work with, ultimately reducing the frequency of errors. In particular, in the case of medical equipment, the design of the user interface can impact enormously on its successful use. In this chapter we consider some of the elements of good and bad medical equipment design, using examples drawn from the literature and elsewhere. The concept of ecological interface design is also discussed, and some practical design guidelines are provided.

INTRODUCTION

American statistics suggest that preventable medical error is the eighth leading cause of death, being responsible for between 44,000 and 98,000 deaths annually, and resulting in injuries that cost between $17 billion and $29 billion annually (Kohn, Corrigan, & Donaldson, 1999). Experts have often stated that an important approach to improving patient safety is to apply system design principles from human-factors engineering (ergonomics; Kohn et al.; Leape, 1994). Human-factors engineering is a relatively young scientific discipline that focuses on those factors that affect the performance of individuals using systems or equipment (Kroemer, 2001). The product may be as simple as a spoon or an office chair, or as complex as an aircraft carrier, but in all cases the goal is to design products to conform to human nature rather than merely expect people to adapt to technology. By doing so, systems and equipment become easier for people to work with, ultimately reducing the frequency of errors.
In the case of medical equipment, the design can impact enormously on its successful use. In particular, errors in operating such equipment are often caused, at least in part, by the design of the user interface. Of course, such errors can not only hamper patient care, but in some cases can even lead to injury or death. It is obviously important that medical equipment be designed with special consideration given the impact of design on safe operation. Thus, the user interface for medical equipment should be straightforward and intuitive: If its operation is excessively complex or counterintuitive, safety can be compromised. Human-factors techniques have been applied to other industries, such as nuclear power and aviation, and have been very successful in reducing error and improving safety in these contexts. Note also that in addition to increasing safety, an added benefit of using good ergonomic design practices is the likelihood that training costs will be reduced.

**BAD DESIGN EXAMPLES**

Examples of perplexing, arcane, and hazardous designs produced in violation of ergonomic principles are not hard to find. For instance, Michael J. Darnell’s Web site www.baddesigns.com offers a collection of frequently humorous examples. But when bad designs in medical equipment lead to injury or death, the situation can be far from amusing. This is sometimes the case for computerized medical equipment.

In one case reported on the U.S. Food and Drug Administration Web site (http://www.fda.gov), a patient was overdosed after a nurse read the number 7 as a 1 in the drug-infusion pump display. Because the flow-rate display was recessed, the top of the 7 was blocked from view at many viewing angles.

In another case report from the same source, a physician treating a patient with oxygen set the flow-control knob between 1 and 2 liters per minute, not realizing that the scale numbers represented discrete, rather than continuous, settings. Unbeknownst to the physician, there was no oxygen flow between the settings, even though the knob rotated smoothly, implying that intermediate settings were available. The patient, an infant, became hypoxic before the error was discovered. One solution could have been a rotary control that snaps into discrete settings.

In yet another case, a patient on a ventilator died following the accidental detachment of the ventilator tubing from the humidifier. Unfortunately, an alarm did not sound because the pressure-limit alarm setting had been set so low that it was essentially nonfunctional.

Finally, Figure 1 illustrates a less hazardous example drawn from personal experience.

A series of reports from the laboratory of Dr. Kim Vicente of the University of Toronto have looked at user-interface issues for patient-controlled analgesia (PCA) equipment (Lin, Isla, Doniz, Harkness, Vicente, & Doyle, 1998; Lin, Vicente, & Doyle, 2001; Vicente, Kada-Bekhaled, Hillel, Cassano, & Orser, 2003). PCA is a computer-based medical technology used to treat severe pain via the self-administration of analgesic agents such as morphine. Potential benefits include superior pain control, automatic documentation, and improved utilization of nursing resources.

Unfortunately, however, one of these units (Abbott Lifecare 4100 PCA Plus II) has been linked to a number of overdose deaths. This machine is easily set up incorrectly by caregivers, who must manually enter the PCA parameters, and a number of patients have received drug overdoses as a result of user errors when using this product: the insertion of a 5 mg/mL morphine cartridge when the machine is expecting a 1 mg/mL concentration, or the acceptance of the default (initial) drug concentration when the correct action is to scroll up to the correct value, among other errors.

In the latter case, when nurses program the drug concentration, the Lifecare 4100 display shows a
Ergonomic User Interface Design in Computerized Medical Equipment

Figure 1. The user interface for medical equipment should be straightforward, friendly, and intuitive. Also, rarely is the operating manual available to the end user, which makes the labeling of the controls especially important. Consider then the user controls shown for this operating-room table in use at the author’s facility. The top and bottom left-hand controls lower the head and feet, respectively, while the right-hand controls raise the head and feet. But what if the entire operating table is to be raised or lowered, which is by far the most common request from the surgeon? It turns out that the entire table is raised by pushing both right-hand buttons, while the entire table is lowered by pushing both left-hand buttons. This arrangement makes sense if one thinks about it for a while, but an intuitive interface should not require a lot of thinking. Furthermore, there is plenty of space available on the control panel to add two extra buttons.

particular concentration (e.g., 0.1 mg/mL). Nurses can either accept this initially displayed value or modify it using the arrow controls. The critical flaw in the design is that in this situation, the Lifecare 4100 offers the minimal drug concentration as the initial choice. If nurses mistakenly accept the initially displayed minimal value (e.g., 0.1 mg/mL) instead of changing it to the correct (and higher) value (e.g., 2.0 mg/mL), the machine will run as if the drug is less concentrated than it really is. As a result, it will pump more liquid, and thus more of the drug, into the patient than is desired.

Aware of the dangers of the Lifecare 4100, Lin, Isla, et al. (1998) and Lin, Vicente, et al. (2001) studied the unit using cognitive task-analysis techniques. Based on this analysis, the interface was then redesigned “to include a dialog structure with fewer steps, a dialog overview showing the user’s location in the programming sequence, better command feedback, easier error recovery, and clearer labels and messages” (Lin, Isla, et al., 1998, p. 253). Studies of the new interface showed significantly faster programming times, lower mental workload, and fewer errors compared to the manufacturer’s original interface. Regrettably, the improved interface design was not used by the manufacturer.
DESIGN GUIDELINES

The U.S. Food and Drug Administration has offered a number of guidelines to help with the design of medical equipment, such as the following (adapted from http://www.fda.gov):

- Make the design consistent with user expectations; both the user’s prior experience with medical devices and well-established conventions are important.
- Design workstations, controls, and displays around the basic capabilities of the user, such as strength, dexterity, memory, reach, vision, and hearing.
- Design well-organized and uncluttered control and display arrangements. Keys, switches, and control knobs should be sufficiently apart for easy manipulation and placed in a way that reduces the chance of inadvertent activation.
- Ensure that the association between controls and displays is obvious. This facilitates proper identification and reduces the user’s memory load.
- Ensure that the intensity and pitch of auditory signals and the brightness of visual signals allow them to be perceived by users working under real-world conditions.
- Make labels and displays so that they can be easily read from typical viewing angles and distances.
- Use color and shape coding to facilitate the identification of controls and displays. Colors and codes should not conflict with industry conventions.

ECOLOGICAL INTERFACE DESIGN

Ecological interface design (EID) is a conceptual framework developed by Vicente and Rasmussen (1990, 1992) for designing human-machine interfaces for complex systems such as are often found in process-control applications or in computer-based medical equipment. The primary goal of EID is to aid operators, especially knowledge workers, in handling novel or unusual situations. Studies suggest that the use of EID methods can improve operator performance when compared with classical design approaches (Vicente, 2002). The basis for EID is Rasmussen’s (1986) skills-rules-knowledge model of cognitive control, and in this model, critical incidents can result from errors at any of the skills-rules-knowledge levels of human cognition.

The first form of incident includes skill-based errors involving the faulty execution of an otherwise correct plan. Here, behavior is unconscious, nonverbal, and automatic. An example would be inadvertently turning on the wrong switch. Even the most experienced clinicians are prone to skill-based errors as they often occur during highly routine procedures such as in reading a drug label or adjusting a control.

A second category is rule-based errors and involves the failure to apply a rule, such as stopping at a stop sign when driving a car, or not administering a drug to which the patient is allergic. At this level, one step up in Rasmussen’s (1986) cognitive model, people use stored (or precompiled) rules acquired with training or with experience on the job.

Lastly, there are knowledge-based errors in which the initial intention is itself wrong, often due to inadequate knowledge or experience. A clinical example would be administering 10 mg of morphine to an infant. This corresponds to Rasmussen’s (1986) highest cognitive level, and is most suited for unfamiliar environments where prior experience is unavailable to provide a system of rules, such as troubleshooting a new piece of medical equipment for the first time.

Under Rasmussen’s (1986) skills-rules-knowledge model, human behavior moves along a ladder as experience increases. Early on, when one is placed in an unfamiliar environment, problem-solving behavior will be at the knowledge level.
As experience is gained so that rules can be formed, the rules level takes over. In some situations, further experience may lead to even further automation (skills level).

For each of the three cognitive levels, the way in which information and environmental cues are perceived differs. Signals guide skill-based behavior, while symbols apply to knowledge-based behavior. Signals supply time-space information only; they have no meaning at higher levels, and they cannot be verbalized. Signs may trigger rules (stop, start, etc.) or may indicate the state of the system (valve open or closed), but they do not express functional relationships (e.g., the consequences of an open valve). Finally, symbols refer to concepts that support analytical reasoning, such as modeling the system to allow one to determine the consequences of an open valve.

What does all this have to do with avoiding human error when operating complex medical equipment? The answer lies in the following. Rasmussen’s (1986) three levels of cognition can be grouped into two broader categories: (a) analytical-based behavior (knowledge-based behavior) and (b) perceptual-based behavior (rule and skill based). Such a categorization is helpful because perceptual processing has important advantages over analytical-based behavior: Analytical behavior is slow, demanding, and serial in nature, whereas perceptual behavior is fast, effortless, parallel, and less error prone. Thus, the goal of design should be to help people avoid situations requiring them to work at the knowledge-based level, while supporting the use of analytical problem solving for use in unfamiliar situations. And, as emphasized above, design guidelines that match the environment to the people involved is known as ecological interface design.

REFERENCES


**KEY TERMS**

**Cognitive Task Analysis (CTA):** A family of methods and tools for understanding the mental processes central to observable behavior, especially those cognitive processes fundamental to task performance in complex settings. Methods used in CTA may include knowledge elicitation (the process of obtaining information through in-depth interviews and by other means) and knowledge representation (the process of concisely displaying data, depicting relationships, etc.).

**Ecological Interface Design (EID):** A conceptual framework for designing human-machine interfaces for complex systems such as computer-based medical equipment. The primary goal of EID is to aid operators, especially knowledge workers, in handling novel or unanticipated situations.

**Ergonomics:** A synonym for human-factors engineering, especially in the European literature.

**Human-Factors Engineering (HFE):** The branch of engineering devoted to the study of the interactions between humans and systems, especially complex systems, with the goal of designing systems that are safe, comfortable, effective, and easy to use.

**Patient-Controlled Analgesia (PCA):** A computer-based medical technology used to treat severe pain via the self-administration of analgesic agents such as morphine.

**Skills-Rules-Knowledge Model:** A multi-level model of cognitive control developed by Rasmussen that is especially helpful in examining human error and critical incidents.

**User Interface:** The components of a system that the operator uses to interact with that system. For example, in the case of a computer, the operator interacts with it using the monitor, keyboard, mouse, and so forth.
Chapter 2.9
Development Of E-Government Services For Cultural Heritage: Examining the Key Dimensions

Carugati Andrea
Catholic University of Lille, France

Hadzilias Elias
Catholic University of Lille, France

ABSTRACT
This article is aimed at defining a framework for the design of e-government services on cultural heritage. Starting from an analysis of three cases on digitization of different types of cultural objects, we highlight the problems existing in the creation of e-services on cultural heritage. These cases show the existence of four key issues in the development of this kind of information systems: digitization, requirement engineering, standardization, and interoperability. The proposed framework addresses these issues, focusing on the user requirements on one side, and the cultural object representation—which is the key to interoperability—on the other. In the cultural domain, the EU Lisbon strategy pushes for the compatibility of shared content across multiple, locally generated contents. Dynamic content exchange requires the use of a prescriptive framework for the development of cultural heritage Web sites. This article provides such a framework, using observation from concrete applications, knowledge of information systems development methodologies, and the IDEF0 modelling method.

INTRODUCTION
The Lisbon strategy for eEurope (EU Report, 2002) and the following eEurope 2002, eEurope 2005, eEurope 2010, drafted as results of the activities of the European Council, are aimed to “make the European Union the most competitive and dynamic knowledge-based economy with improved economy and social cohesion by 2010.” In concrete terms, this means broadband and high-level Internet-based services for the entire population of the European Union. The means envisioned to achieve this goal are largely based
Development Of E-Government Services For Cultural Heritage

on increasing both demand and offer of e-services respectively, from the public/users and the providers. The problem has been framed as a “chicken and egg” problem and the solution has therefore been to address both ends: increase government-side services and create a friendly legislation for the implementation and sale of broadband connections (EU Report, 2002). This article focuses on the demand side, that is, on the development of the public electronic services.

On the demand side, electronic government initiatives involve providing services in e-government, e-learning, e-health, and e-business (EU Report, 2002). While the efforts of e-government are focusing on providing services to citizens in order to achieve higher efficiencies through automation (tax filing, certification, electronic voting, information provision, and so forth), one other important area of investment regards the facilitation of access to cultural resources. The regional and local cultural heritage (to be defined in a broad sense, from museums to regional gastronomy and folklore) is one of Europe’s greatest economic assets, and ICT and other advanced technologies can dramatically increase the possibility of its exploitation. Until now, the government initiatives for divulging electronic material on the local cultural heritage have been varied in nature and include the creation of portals for information on cultural events, which is the most common model of exploitation today, including the digitization of artwork for archives, the creation of virtual tri-dimensional museum visits with tri-dimensional digitization of art works, and the rendering of archaeological visits in digital formats (Carugati, Hadzilias, & Demoulin, 2005).

Nevertheless, the potential of using electronic services for cultural heritage applications is far from being fully exploited and many initiatives have remained at the stage of pilot projects. Of these pilot projects, few are completed and most contain only one or very few examples of art digitization. Until now, experiences of use of ICT in cultural heritage sectors too often fail in provid-

ing valuable economic results due to a number of problems, and have generated disappointment among the potential players and beneficiaries. The main problems have been:

• In general, at regional and local level, there is shortage of experience and expertise about the use of ICT in cultural heritage areas. Therefore, local and regional administrators have to rely on ICT vendors and consultants, and these in general, are mainly able to suggest general purpose solutions, non-optimised for the specific sector, since even large IT consulting companies have limited expertise in the cultural heritage area.

• If we consider the “conventional” cultural heritage, such as museums and galleries, this sector lacks expertise and experience in marketing and business promotion on electronic media, which makes it difficult to develop credible business models and plans, and to attract investments.

• There are analogous problems also in the cultural tourism sector. There are hundreds of projects and initiatives related to cultural tourism in Europe, but they often have been designed on “supply oriented” thinking, without systematic investigation into what the customer, the “cultural tourist,” is looking for. The issue of user diversity should be considered in the system development process, as pointed out by Klawe and Shneiderman (2005), since it is a critical success factor to steer user preferences. This is valid not only for business purposes, but also and foremost at the community level (Carroll, 2001). Finally, as e-government services are web-based, they are available to different users. Even though the expectations, in terms of services, might differ, Nielsen (2005) states that usability factors influence, for both professionals and other categories, the success of the service.
More generally, the problem of mutual understanding between the cultural sector, service industry, and ICT experts, due to their different ways of thinking, makes it difficult to achieve a fruitful cooperation in areas ranging between education and multimedia production, where the availability of content from the cultural heritage could open very promising development perspectives. In order to be able to develop innovative and marketable products and services with a real cultural added value, the personal direct involvement of experts from the cultural area is absolutely required.

Despite the above-mentioned issues, in Europe as well as elsewhere, there are a number of interesting examples and initiatives, on various scales, of successful economic promotion of the cultural heritage (some of these mentioned above). Unfortunately, they have often only a local or regional visibility, and their positive (and negative) experiences cannot be fully exploited and shared by other projects.

To resolve these issues, the article proposes an integrated framework for developing e-government services on cultural heritage. The framework emerges from the study of multiple cultural heritage electronic initiatives and the in-depth investigation of three case studies. This framework is represented using the activity modelling method IDEF0, where all necessary steps, inputs, outputs, rules, and roles are described in a hierarchical manner.

The article is organized in the following way: first, the research methodology is explained, secondly, we present the case studies and analyse them to highlight key issues, we describe the IDEF0 modelling technique, and then we continue presenting the development framework, focusing on user segmentation and interoperability. Finally, we conclude with a discussion of the proposed framework, highlighting the need for a systemic approach to the activities.

**RESEARCH METHODOLOGY**

The research was carried out following the constructive paradigm of the case study approach (Yin, 2002). In this context, the research used a practice-based lens to identify processes and problems in the development of e-government services on cultural heritage. The process was comprised of three steps:

1. Identify representative cases of e-government cultural services;
2. Identify the different functionalities for the expected users and the problems incurred in these specific cases of information systems development;
3. Propose a way to consider these problems given the constraint of limited resources.

As a first step, we carried out an extensive survey to record most electronic initiatives in the culture domain. The survey identified 142 initiatives, which should cover the bulk of such services for the eight languages spoken by the authors. The languages cover the vast majority of projects in this domain. Among these initiatives, we have selected three that are the most representative because they fit specific criteria that we set out in our research: 1) they continue to exist after the end of the pilot project; 2) they are complete in terms of the cultural content they were supposed to cover; 3) they provide additional functionalities with respect to similar services; 4) they use more advanced technologies; and 5) the informative material about the case was comprehensive.

As a second step, we analyse the three cases in a structured manner looking at a) functionalities provided to the users; b) functionalities provided to other institutions; and c) the development process. These three elements are compared and integrated to indicate key leverage points of these initiatives. In this step, we include usability testing for the two initiatives available online (the
Piero della Francesca Project and the incunabula manuscripts in the Herzog August Library) in order to see if the users find the functions of the Web site friendly and value-adding. For doing this, we carry out the testing with five users, representing different categories (Nielsen and Landauer, 1993). The layout of the testing follows the indication of Molich (1999, p. 109), with a testing room and an observation room. According to the testing protocol, the users were given different tasks to carry out and were required to speak over their actions.

In the third step, we employ the IDEF0 modelling method to describe a high level design framework for implementing e-government services for cultural heritage. The framework expands on treating the key points identified in step two. The proposed solution is therefore the result of deductive reasoning and is prescriptive in nature.

CASE STUDIES

According to the selection process, we consider the following case studies: the Electronic Beowulf Project at the British Library (UK), the Piero della Francesca Project in Arezzo (Italy), and the digitization of the incunabula manuscripts in the Herzog August Library in Cologne (Germany).

Beowulf: The Pioneer Project

The Electronic Beowulf Project represents a pioneering effort in the digitization of cultural heritage. Beowulf is both the first English literary masterpiece and one of the earliest European epics written in the vernacular, or native language, instead of literary Latin. The story survives in one fragile manuscript copied by two scribes near the end of the 10th or the first quarter of the 11th century. Until quite recently, most scholars thought that this surprisingly complex and poignant poem was written in the 8th century or earlier, but Kiernan (1981) stirred up controversy asserting that the work was composed in the 11th century, and that the manuscript itself may have even been the author’s working copy. The manuscript has been damaged over time and the original that is surviving today is in very poor condition (“Why is Beowulf important?”, 2005).

The Electronic Beowulf Project was carried out in the period 1993-1999 in order to digitize the original Beowulf manuscript, of which the original is now in the British Library, and has been made accessible to everyone in electronic format.

Development Process

Since the Beowulf represents one of the first attempts to digitize a cultural heritage object, many difficulties were encountered in different aspects of the project, as the work in preparing the digital edition was very complex in the period considered. Technical difficulties concerned scanning technologies, storing media, and managing the transfer across large distances of large quantities of data or moving 24-bit colour images across different technical platforms. In some cases, it was necessary to devise innovative technical solutions to achieve the desired end result. The following is a quote from Prof. Kiernan’s Web site, highlighting the technical difficulties encountered in 1993:

The equipment we are using to capture the images is the Roche/Kontron ProgRes 3012 digital camera, which can scan any text, from a letter or a word to an entire page, at 2000 x 3000 pixels in 24-bit color. The resulting images at this maximum resolution are enormous, about 21-25 MB, and tax the capabilities of the biggest machines. Three or four images—three or four letters or words if that is what we are scanning—will fill up an 88 MB hard disk, and we have found that no single image of this size can be processed in real time without at least 64 MB of RAM. In our first experiments in June with the camera and its dedicated hardware, we transmitted a half-dozen...
images by phone line from the Conservation Studio of the British Library to the Wenner Gren Imaging Laboratory at the University of Kentucky, where identical hardware was set up to receive the data. Most of these images are now available on the Internet through anonymous ftp or Mosaic ("Electronic Beowulf: A Guide," n.d.)

Besides the technical challenges, there were also issues in the coordination and collaboration of the human resources, due to the variety of disciplines necessary to carry out this task. The domain experts comprised scholars, curators, conservators, photographers, and technical experts (Prescott, 1997), who had to agree on how the activities had to be carried out in order to record the maximum information without damaging the precious document.

In the case of Beowulf, the development was mainly oriented towards the utilization of information technology for capturing cultural content. This activity was not only done to preserve the artifact, but also to enhance the comprehension of the text. From the multiple sources about this project, we can see that, given these requirements, the main activity was to create digital images of the text with different scanning techniques (B&W, 24 bit color, ultraviolet, and high level fiber optic light). This was performed once in order to minimize damage and retain the maximum information digitized because of the extensive erasure and correction in the original manuscript.

The development process followed was a pure waterfall approach (Boehm, 1988). In the situation, the requirements had to be set at the offset of the project and the digitization was done only once per page. Afterwards the content was used to populate a database.

Functionalities for Users

The Electronic Beowulf Project is available to users on CD format that can be purchased from the British Library. The interface provides access to the digital content of a database that contains both pictures and transcribed text. The CD contains multiple transcription of the Beowulf from different accredited authors. Different search criteria are available to query Line (Edition), Folio (Edition), FolioLine (Transcript), Fitt (Edition or Transcript), Scribe (Edition or Transcript). Images can be zoomed to a very high detail in order to be studied by paleographers and calligraphers (see Figure 1). Alternatively, the layout of the screen can show both the original manuscript and the transcription, which is intended to be used by researchers interested in the content of the poem (see Figure 2).

The electronic version enables readers to place the original manuscript’s leaves side by side, to examine the color and texture of the vellum leaves by magnifying the images and to explore the work of the early scribes (see Figure 3). Building on this material, the CD features multiple versions of restoration of which the most recent one was made possible by the different scanning techniques.

Functionalities for Institutions

This project was part of the strategy of the British Library to create the capability to increase access to its collections by use of imaging and network technology. As such, the Beowulf project was not directly geared towards providing content, services, or functionalities to other institutions and remains a stand alone product.

One of the collaborations that emerged from this project was the inclusion of the Thorkelin’s transcript of the Beowulf catalogued in the Royal Library of Copenhagen.

Piero della Francesca: From Restorers to Children

The project of Piero della Francesca regards the digitization of the series of frescos “Leggenda della Vera Croce” situated in the Bacci chapel in the Basilica of St. Francis of Arezzo (Italy). The project was started in the mid 80’s to repair the
Development Of E-Government Services For Cultural Heritage

severe damages that the frescos suffered in the early 80’s. The project was lead by the cultural and artistic heritage authority of Arezzo (Soprintendenza per i Beni Architettonici, Artistici e Storici di Arezzo). The causes of the damage were not clear and therefore the Soprintendenza decided to begin a series of deep and comprehensive scientific and historic research studies to determine the reasons for the fresco’s detrimental conditions. The aim was the preservation of the masterpiece for future generations, and in order to achieve this, it was necessary to find the right technique to restore the painted plaster as well as the building structure itself.

As the project evolved and the restoration work finished, the initial goal was complemented by the creation of a permanent documentation center “for future study and diffusion of information regarding the study and conservation of the mural paintings of Piero della Francesca and to construct, with the use of computer technology, an organic database for the Project and an interactive multimedia system to be hooked up to the international electronic network and available for public consultation.” The project is considered one of the most advanced and demanding efforts of the entire scientific restoration field in Italy and abroad.

Development Process

Originally, this project was carried out to provide professional restorers with new tools for restoring the artworks of Piero della Francesca. For this reason, the initial requirements were set by the restorers and as a result, the quality and detail of the digitized material is very high. For example, they integrated text, vectorial information (CAD
drawing), and raster information (images) which, from a data processing point of view, are of a completely different nature and very complicated. Figure 4 shows the process followed by the project team from preliminary study, to material collection and data processing, and finally the development of software applications.

The four basic activities in this project: data capture, CAD drawing, and raster images (overlaid) are presented in Figure 5.

As in the Beowulf case, the development approach followed the waterfall model. Restorers, together with state institutions, and private and public research institutes, carried out a long preliminary research to set out the requirements. Leading scholars and expert technicians used state of the art information technology to develop the applications that were used to successfully define the restoration approaches to solve the fresco complex problems.

After the end of the restoration work the development continued with the creation of a Web site that used the same collected material from the professional applications, but provided functionalities that were interesting for the wider public. In this second cycle of development, a new set of requirements was created to address the goal of creating a permanent documentation center.

Functionalities for Users

For the restorers, the developed applications provided functionalities to store and search thousands of documents; illustrations of the church choir were compiled in a basic cartographic model and in a variety of thematic mappings with reference to historical studies, visual surveys, diagnostic studies, and scientific analyses and complemented by about 1000 historical/chronological records. The catalogued images could be called up individually from a special menu where they were arranged according to distinct categories. This technology was particularly effective in the phases of correlation, control, and updating of the various studies.
conducted. The images could also be called up from a relational database, which allowed the efficient consultation of records.

Furthermore, the software allowed keeping record of the restoration in process (work site diary, graphs and charts of the work performed, technical specifications, and so forth) as well as adding new cultural content.

To help restorers in their work, the applications operate in the complete respect of the iconographic and technical value of the artist’s work, providing a documentation of the various painting techniques employed by the artist: from the transfer of the preparatory design through the use of stencilled cartoons, etched lines and colored threads, to the application of color to the fresh plaster (a fresco), partially dried plaster (a mezzo fresco), and finally to the dried plaster with special temperas and organic blending agents. A further example of the system’s capabilities is that it provides a chronological reconstruction of the artist’s production. Connected to this, the system allows the study of the relationship between the saline deposits and the presence of fixatives added during earlier restorations or, alternatively, to view the deposits in relation to the various painting techniques used by the artist.

To satisfy the need for rapid consultation in the service of monitoring the state of conservation and the progress of the restoration, system designers developed a thematic glossary with a unified terminology of key words for research and access to the data, organized first in separate technical files and then electronically translated to permit interactive consultation.

Finally, the applications were online with the work site, in order to manage the traditional paper documentation kept by the project director and the restoration technicians and allowing them to have immediate access to and use of the data stored in the system.

In the public Web site, one of the services provided is the classic function of archive. The archive allows searching for historical events by date or by key word and presents the results combining digital images with metadata. The result of one search is presented in Figure 6.

Other interesting functionalities provided are games based on the thematic mappings. The Web site proposes two games: one that consists of adding colours to a black and white painting and the other where parts of the paintings have been divided (electronically) into pieces that the player is supposed to reassemble like a puzzle.

Figure 6. Public archive from the Web site http://www.pierodellafrancesca.it
These two games have an underlying pedagogic reason and would not be possible without the existence of the digitized material specifically created and mapped for the restorers. The Web technologies deployed for this purpose were active images and java applets. An example of the two games is presented in Figures 7 and 8.

**Usability Testing**

The usability testing was carried out by five people: two laymen, one professional restorer, and two children aged seven and eight years old. The adults got the task of finding the painting of “La Madonna del Parto,” starting from the home page. The children had the task of solving the two games. All subjects reported poor navigation between the different pages due to non-existent text explanation of the links and unconventional graphics and structure. Both the laymen and the professional restorer found the painting in approximately three minutes and using 30/32 mouse clicks (against one minute and eight clicks that it would take for an experienced user of the system). The information presented pleased the professional, but not the laymen that reported it as too complex and dull. All three cases showed a critical design error when the user had to pass from the classic Web site view to the database of the art works, because we had to intervene in order to guide the users on how to continue their search.

In consideration of their age, the children were not asked to navigate through the Web site and were presented directly with the games. The

**Figure 7. Puzzle game**

![Puzzle game](image)

**Figure 8. Painting game**

![Painting game](image)
children did not understand what they had to do and also complained that there was no evaluation of their performance. Using the words of the children, these games were reported to be: boring and meaningless.

Despite the interesting use of the digitized material for a wider audience, the outcomes were not value-adding, neither for the laymen nor for the children.

Functionalities for Institutions

The system provides two levels of access to the project material. At the first level, the information is available on the Web site that is hosted by the sponsor of the project (Banca Etruria) and at the second level, which is more complete and contains the full record of the restoration project, the system is available at the offices of the cultural and artistic heritage authority of Arezzo. This local system can be accessed only under official authorization. Data can be processed through a cartographic system based on photogrammetric surveys of the painted walls. The relational database is structured in a series of indexes. Files are divided into didactic, diagnostic, and intervention sections. The local system is addressed to institutions that seek in depth knowledge, not only of the artworks, but also of the restoration process, the mineralogical analysis, and the chemical analysis performed. The left side of Figure 9 shows the local cartographic system and the right frame shows the functionalities for professionals. The green dot indicates the access to the public archive via the Web.

Distributed Digital Incunabula Library: An Architecture for Sharing

The “Distributed Digital Incunabula Library” (inkunabeln.ub.uni-koeln.de/vdib/) is a project aimed at the digitization of all German incunabula. Incunabula, Latin for cradle, are books printed between 1454 and 1500. Worldwide, approximately 550,000 incunabula have been preserved. The Distributed Digital Incunabula Library project was organized as a joint effort of the “Koelner Universitaets und Stadtbibliothek” and the “Herzog August Bibliothek Wolfenbüttel” libraries. The two libraries involved, hold together approximately 5,800 copies incunabula. Out of these, 350,000 pages (approximately 2,000–3,000 titles) were selected for this project, of which 40% were digitized in the first ten months of the two year effort. When completed, the 350,000 pages could represent as much as six percent of the titles of incunabula preserved world wide (Loebbekke and Thaller, 2005).

As a means of achieving global integration while simultaneously keeping institutional in-
dependence, this project proposed as reference architecture a new tool called “Digital Autonomous Cultural Objects (DACOs).” DACOs are characterized by a common behavioral code that provides information about cultural objects on request. The code involves the object wrapping itself in mark-up language. It also provides a persistent addressing scheme, allowing for control of the individual institution holding the good as well as a mapping scheme. Thus, even if the technology (e.g., HTML, XML, SGML, VRML objects, Flash movies, SVG constructions) can be chosen by the individual institution, the structure of basic elements integrated into a DACO is predetermined.

Development Process

In order to deal with the vast amount information to be digitized, a standard development process was followed, which is depicted in Figure 10.

The first stage of the development was the raw digitization of the single pages, which was conducted using a Nikon DXM 1200 and a camera server that stored the images with total capacity of 80 GB. A book cradle for photos in an angle of 45 degrees (see Figure 11) allowed treating the incunabula carefully while simultaneously saving time and money. The pages were scanned in a 24-bit color scheme using a resolution of 3800x3072 pixel leading to a data file size of 34 MB/page. With such a digitization procedure, more than 80% of incunabula were digitized leading to about 12 TB raw data volume.

At the second step, the digitized data was transferred through a local area network to a pre-processing server with 250 GB storage capacity. This server was responsible for the transformation of the output files to the size that is efficient for Web site access and for the different users.

Finally, the files were forwarded to the Web server environment and the digitized files were available on the Internet, after passing a quality control at the Metadata Cache and being transformed to DACOs. These versions of the original files were not more than 10% of the original data size, so that they can be accessed quickly from the Internet. During this phase, the data supervision and the production of the copy-referred development data took place. The archive server offers a storage area for disaster recovery purposes.

The entire architecture was based on cost-efficient Linux servers highly performing for the hierarchical storage organization.

Obviously, beyond tackling the technical challenges, cost played a major role in pushing such
Development Of E-Government Services For Cultural Heritage

...a digitization effort forward. The most important cost factors were:

- Content and subject classification based on the already existing Incunabula Short Title Catalogue (ISTC);
- Workflow/data importation;
- Raw digitization;
- Storage capacity;
- Development and maintenance of the WWW server.

Considering all cost components in this project, the costs for one page amount to roughly 0.75 € (plus metadata and work environment) (Loebbecke and Thaller, 2005).

Functionalities for Users

The users are able to run queries in the incunabula resources in two levels: text-based and image-based. In the former, there are query criteria for locating the page or pages under consideration, containing all metadata and the digitized images. In the latter, there is an image archive with three sublevels of incunabula description: the Incunabula Short Title Catalogue (ISTC) classification, the original image, and the transcription of the image. Furthermore, additional tools are provided allowing for browsing through the manuscripts like in traditional libraries, as shown in Figure 12. A native “dynamic” XML database administers all codicological manuscript descriptions. For some codices, those descriptions amount to the equivalent of fifty printed pages.

There are four different viewing levels of the scanned images suitable for the categories of users who will study the cultural object. Figure 12 shows the query result with the overview level (notice the eye, the glasses, the loop, and the microscope icons above the scanned image). Figure 13 shows a higher level of detail.

Another e-service offered is that the users can download several tools, which measure the performance of typical types of codicological/palaeographic work (e.g., measurement of single characters). As far as available without copyright restrictions, the main scientific literature to work with the codices is also made available in digital form.

Using DACOs, one URL of the library dissolves into about 600,000 URLs of individual...

Figure 12. Query results (http://inkunabeln.ub.uni-koeln.de/ visited on 25/7/05)
references to a specific cultural heritage object that can be referenced directly from any other cultural heritage resource (see, for example, the code \texttt{ig00227000} above the scanned page in Figure 12). This is due to the various views of the metadata and different resolutions.

**Usability Testing**

Given the content of this site, we had to carry out a careful selection of the five users. Finally, we selected three librarians and two history researchers. Even though the test subjects did not research specifically the incunabula, they were selected for their knowledge in the research and management of historical documents. They were given two tasks: to find the document “Mensa philosophica” and the complete production of Gerson. The users completed the first task in less than two minutes and with 10/14 mouse clicks (compared to an experienced user that requires seven mouse clicks). They completed the second task, restarting from the main page, with seven mouse clicks in less than one minute. The users reported the site to be easy to navigate, even though it was in German language. However, they evidenced two major failures in the database operation: the field tables were not linked and the navigation functions in the index, instead of visualizing the records alphabetically, were providing erratic results.

**Functionalities for Institutions**

The highest added value of this project lies in the inter-institutional, cultural content sharing. Cultural heritage brokers, connected to servers recognizing DACOs, integrate many DACOs from different institutions into specialized interfaces, thus serving the interests of laymen and researchers. The communication between the servers takes place via the DACO protocol consisting of several XML-codes transmitted via Hyper Text Transfer Protocol (HTTP). When addressed by a broker, the DACO server describes itself through providing a list of supported “access venues” (e.g., titles, authors, years, and so forth). The access venues can be used by the broker to develop a search mask.

**CASE ANALYSIS**

After examining thoroughly the three case studies of e-services in the cultural sector and with the knowledge of the other existing initiatives, we observed that the greatest challenge lies on the management of four dimensions: digitization,
Development Of E-Government Services For Cultural Heritage

requirement engineering, standardization, and interoperability.

The survey we carried out shows that several European countries are making efforts to comply with the Lisbon strategy and are moving, albeit at different paces, towards the digitization of their cultural heritages. However, given the sheer size of the material, the variety of the applications developed, and approaches taken by the different governments, it appears clear that the seamless integration of the cultural heritages of multiple countries is a very complex task and it is addressed in multiple ways. The three cases showed that digitization is the starting point of these efforts, aiming at the preservation of the cultural content and its availability to a large spectrum of users. All three cases show a very similar approach to the digitization problem, using various scanning technologies to create digital facsimile of cultural objects.

The second dimension concerns the requirement engineering. In the three cases observed, the requirements were set initially by the most demanding groups, namely the professionals working with cultural objects in their daily working life. Researchers were in contact with the technical developers in order to discuss and define the initial set of requirements. This helped to set very high quality standard for what later became material available to laymen. The process of creating the Web sites available to the public was carried out following a top down approach that provided the digitized content to the public without examining the specific needs of the different categories of users. A notable exception is the functionality offered in the Piero della Francesca project where the pedagogic games addressed the needs of the specific target group of children. However, this effort followed the same push strategy as the other two and our usability tests proved the limitation of this approach.

As far as standardization is concerned, there seems to be a trend towards finding a common way

Table 1. Overview of the findings

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements engineering</td>
<td>Top-down approach from professionals’ perspective</td>
<td>Top-down approach from professionals’ perspective</td>
<td>Top-down approach from professionals’ perspective with special focus on interoperability.</td>
</tr>
<tr>
<td>Interoperability</td>
<td>Not addressed. Available on CD-ROM.</td>
<td>The archive is partially available online, but the content is not accessible from other Web sites.</td>
<td>Content shared with other institutions thanks to specially designed XML objects (DACO).</td>
</tr>
</tbody>
</table>
to represent cultural objects. In the early cases, the standardization regarded the format of the digitized images (e.g., jpg, tiff, bmp formats). Instead, the incunabula project shows a more structured approach to the standardization problem, using the DACOs where, not only the images, but also the metadata, follow a specific standard.

Finally, we have also observed an evolution towards higher interoperability of cultural information systems. The early example of the Beowulf was designed as a stand alone project and it is only available to users on CD-ROM to be bought from the British Library. The Piero della Francesca project started as an independent solution for professionals, but it was later made available on the Web with some of the original functionalities activated. The incunabula project was—from the offset—born as a Web-based information system to execute queries on the pages of digital incunabula. This last effort is directly geared towards making the material and the tools available to other cultural Web sites. In order to increase the interoperability of existing or future applications, there should be distributed, localized development of content that can be integrated as it is created. The summary of the case analysis is shown in Table 1.

In an era where content exchange is becoming the norm, and access to cultural content is mandate by law (eEurope strategy), it becomes paramount for the different governments to develop their content, respecting a framework that will allow different Web sites to freely exchange content. The framework will have to address the four dimensions just described and embed them into a process of information system development. The development of such a framework will be the focus of the next part of the article, beginning with the description of the modelling technique used—IDEF0.

**IDEF0: THE MODELLING TECHNIQUE**

In order to model our framework, we selected a modelling technique that was suitable for easily conveying a message and, at the same time, did not impose methods or techniques on the system modelled. Among the possible techniques (e.g., Data Flow Diagram, ARIS, UML, and so forth), we have selected IDEF0 for its simplicity, well knowing that experts easily translate it into any other modelling tool.

The IDEF0 modelling technique is intended to model the decisions, actions, and activities of a system. It is not only the most widely used, but also the most field proven function modelling method for analysing and communicating the functional perspective of a system (Malhotra and Jayaraman, 1992). IDEF0 was derived from a well-established graphical language, the Structured Analysis and Design Technique—SADT (Marca and McGowan, 1988). The IDEF0 modelling method establishes the scope of analysis either for a particular functional analysis or for future analyses from another perspective. As a communication tool, IDEF0 enhances domain expert involvement and consensus decision-making through simplified graphical devices. As an analysis tool, IDEF0 supports the identification of the functions performed and what is needed to perform them.

The basic activity element of an IDEF0 model diagram is represented by a simple syntax. A verb-based label placed in a box describes each activity. Inputs (I) are shown as arrows entering the left side of the activity box, while the outputs (O) are shown as exiting arrows on the right side of the box. Controls (O) are displayed as arrows entering the top of the box and mechanisms (M) are displayed as arrows entering from the bottom of the box. Inputs, Controls, Outputs and Mechanisms (ICOMs) are all referred to as concept. An IDEF0 model diagram is
Development Of E-Government Services For Cultural Heritage

then composed of several activity boxes and related concepts to capture the overall activity. IDEF0 does not only capture the individual activities, but also reveals the relationships among activities through the activities’ related concepts. For example, the output of one activity may become the input, control, or even a mechanism of another activity within the same model.

A strategy for organising the development of IDEF0 models is the notion of hierarchical decomposition of activities (see Figure 14). A box in an IDEF0 model represents the boundaries of an activity. Inside that box is the breakdown of that activity into smaller activities (child diagrams), which together comprise the box at the higher level (the parent). This hierarchical structure helps the analyst keep the scope of the model within the boundaries represented by the decomposition of the activity. This organisation strategy is also useful for hiding unnecessary complexity from view until a more in-depth understanding is required.

One of the most important features of IDEF0 as a modelling concept is that it gradually introduces

Figure 14. Hierarchical decomposition of an IDEF0 model diagram
greater and greater levels of detail, through the diagram structure comprising the model. In this way, communication is enhanced by providing the reader with a well-bounded topic with a manageable amount of detail to learn from each diagram. The decomposition is graphically indicated with codes; the highest level is called A0, and the activities which are included are indicated with A1, A2, A3, and so forth. (see Figure 14).

An IDEF0 model starts by presenting the whole subject as a single unit—a box with external-arrow boundary conditions connecting it to functions and resources outside the subject. The single box is called the “top box” of the model or the context diagram. Since the single top box of an IDEF0 model represents the subject as a whole, the descriptive name in the box is general. The same is true of the external arrows of the model, since they represent the complete set of external boundary conditions of the subject as a whole, including access to mechanism support that supplies additional means of performance.

Description of the Proposed Framework

Based on the analysis of the cases, the exhaustive survey of the cultural heritage initiatives, and our theoretical and practical knowledge on information systems development, we propose a framework for the development of e-services for cultural heritage based on four ideas:

1. E-services for cultural heritage are, in essence, information systems. As such, existing methodologies can be adopted and adapted for their development.
2. IDEF0 is a proven modelling technique for the development activities of information systems.
3. The impact of Web sites for cultural heritage increases with the enhancement of content integration from multiple sites: a one-stop cultural heritage site can be imagined as the ideal goal and therefore compatibility of the finished local products is paramount.
4. Multiple types of public have to be served by these sites; satisfying each group is important to serve the goal of increasing demand for services in the eEurope strategy.

According to these four ideas and the four dimensions identified in the analysis, we developed a framework which contains the following five phases: collection of user requirements, digitization of cultural content, system interoperability design, cultural content organization, and e-government cultural portal development. These phases are expressed as activities in the IDEF0 model of Figure 15.

The mechanisms that we have identified are:

- Users. This mechanism includes art specialists (like the restorers in the Piero della Francesca case), historians, researchers, and the public in general. In Figure 15, it is shown that the mechanism “users (M1)” supports the collection of users’ requirements (A1), the organization of cultural content (A4), and the development of e-service (A5). The participation of users in these activities ensures the user-centered evaluation that the usability testing proved essential to avoid critical errors.
- Technical project team. It includes the project manager, the system architects, the system analysts, the database designers, and the developers, including the Web designers. In figure 15, it is shown that the mechanism “technical project team (M2)” supports the collection of users’ requirements (A1), the organization of cultural content (A4), and the development of e-service (A5). The participation of the technical project team in these activities ensures that the users’ needs can be supported by the technology and the resources available.
The cultural organization. It includes all the experts in the specific art works that are part of the cultural heritage e-service (curators, conservation scientists, archivists, museum technicians). This mechanism is discrete from the users because they provide information about the art works and are not necessarily using the service. The participation of the cultural organization in these activities ensures that the art work will be treated in the proper way, that the needs of other cultural organizations, in terms of exchange of the digitized materials, are upheld, and that the metadata is correct and complete.

The specific characteristics of each activity are presented in detail in the following sections.

**Collection of User Requirements**

This activity is aimed at the creation of the requirements document for the development of the e-service. To design an efficient cultural heritage e-service that meets the public needs, it is first of all necessary to define the target group users and understand the requirements that have to be addressed. The survey of the cultural heritage initiatives highlighted different groups of users. One classification (see Dipartimento dei Beni Culturali, Italy) identified users as internal, private (customer market), or corporate (business market). For example, the internal public might be the law enforcement agencies accessing the artwork database for research on fakes or stolen material. The differentiation between private and corporate is quite unusual in a governmental service, since every citizen should be equal in front of the state. This might be the indication of a trend to adopt models from e-commerce into
Development Of E-Government Services For Cultural Heritage

e-government, or it might be the result of using technology providers used to dealing with the private sector instead of with the government. The three cases analysed above show that segmentation is deemed important by the creators of cultural heritage e-services.

The input of this activity is a plan for the creation of an e-service for cultural heritage. The plan may originate from a specific need like in the case of Piero della Francesca, from a research interest like in the Beowulf case, or from a long term plan for the distribution of content like in the incunabula project.

User segmentation (Control C1 in Figure 15) guides how the requirements will be created. The users can be classified according to their profession, location, affiliation, age, and interests (Sen, Padmanabhan and Tuzhilin, 1998). Region and language can be used to make sure that the service provided is culturally acceptable and the language is understandable by the audience. Among demographical variables, age may be useful to present cultural heritage accurately at every group age. Occupation or community of practice of the user is a strong determinant of the requirements for the e-service.

As these services are rather new, it is difficult to fathom what would be the requirements of the different groups. Therefore, it is important to develop prototypes for the requirement definition. The user-centered evaluation, following the testing protocols of Nielsen and Landauer (1993) and Molich (1999), should be carried out to assure that the technical project team has well interpreted the user requirements. The testing protocol is represented by the control C2 in Figure 15.

If the information is missing because the service is new than the cases indicate that it is convenient to serve the most demanding group first, like the restorers in the Piero della Francesca case, and then reduce the complexity as other user segments are identified.

The groups involved in this activity are the e-service project team (the technical developers), the cultural organization that owns or manages the art works, and the users in their different declinations. The users are necessary because they will be able to state what the needs are. Users should be representative of the categories identified in the segmentation exercise. The technical team is necessary because of the preparation for the technical work ahead. They have to make sure that the requirements can be technically satisfied. Finally the cultural organization is also necessary because they know how the art works can be handled and what are the aspects of their digitization that are most valuable to cover. As an example, the case of the Beowulf showed that it was necessary to do the scanning under UV light and only specialists could point that out.

Despite a careful segmentation, the actual collection of the requirements can result in a quite challenging process. Most constituencies, especially layman, will in fact be unaware of what they want, since innovative e-services are still very limited. While the mental model of a museum or an archaeological visit is shared among people of different countries, there might be requirements that no user can state, simply because one cannot ask what is not known. For the elicitation of requirements from the final users, it would then seem appropriate to engage with the different constituencies in an iterative process using prototypes of interfaces. This prototype-based requirement elicitation should be done autonomously from the setting of the requirements for the content digitization, where the use of an agreed upon standard is required.

The output of this activity is the requirement document. For the technical project team, it will contain the functional specification of the e-service to be developed. For the cultural organization, it will contain the specification for the preparation of the art works and the specifications for the descriptions to be prepared.
Digitization of Cultural Content

This is the activity of digitization of cultural content. The reasons for digitization, or more precisely for the digital conversion of non-digital cultural source material, are varied and may well overlap. The decision to digitize has the following objectives:

- To increase access: this is the most obvious and primary reason, where a high demand from users is estimated and the administration has the desire to improve public access to a specific collection.
- To improve services to an expanding user’s group by providing enhanced access to the institution’s resources with respect to education and long life learning.
- To reduce the handling and use of fragile or heavily used original material and create a backup copy for endangered material such as brittle books or documents.
- To give the administration opportunities for the development of its technical infrastructure and staff skill capacity.
- To develop collaborative resources, sharing partnerships with other administrations to create virtual collections and increase worldwide access of a country’s heritage.
- To seek partnerships with other administrations to capitalize on the economic advantages of a shared approach.
- To take advantage of financial opportunities, for example the likelihood of securing funding to implement a program, or of a particular project being able to generate significant income.

According to which of these needs are deemed more important, the requirement document is taken from the previous activity and transformed into two outputs: the digital facsimile and a new systems specification document revised after the digitization. The technical project team and the cultural organization collaborate in this activity which, as seen in all three cases, is very demanding.

The control is provided by the MINERVA guidelines. MINERVA, MInisterial NEtwoRk for Valorising Activities in digitization (http://www.minervaeurope.org/), is a resource for guidelines concerning the digitization of cultural content. MINERVA was created by a network of European Union (EU) Member States’ Ministries to discuss, correlate, and harmonize activities carried out in digitization of cultural and scientific content for creating an agreed European common platform, recommendations, and guidelines about digitization, metadata, long-term accessibility, and preservation. Due to the high level of commitment assured by the involvement of EU governments, it aims to coordinate national programs, and its approach is strongly based on the principle of embeddedness in national digitization activities. The use of the MINERVA guidelines insures the contacts with other European countries, international organizations, associations, networks, international and national projects involved in this sector.

The output of this activity is the digitized content. This content can exist independently from the creation of the e-service that uses it (like a repository) or will be used in the development activity. The other output is the revised version of the specification document in accordance to the problems and opportunities that the digitization phase might bring to the surface.

System Interoperability Design

This activity is aimed at assuring the compatibility of the content generated across e-services. This activity takes as input the systems specifications produced during digitization and adjusts them for interoperability.

For this activity, we propose as control the adoption of the ISO 21127 standard (http://www.iso.org/). ISO 21127 is a domain ontology for
cultural heritage information. As such, it is designed to be explanatory and extensible, rather than prescriptive and restrictive. Currently, no specific formalism for semantic models has been widely accepted as standard, which is proven by the existence of the DACO used in the incunabula case. However, the semantic deviations between the various available models are minimal. At this stage, we prefer to recommend the use of ISO 21127 instead of the DACO as control, since the DACOs are not recognised as a standard yet.

The ISO standard has been formulated as an object-oriented semantic model, which can easily be converted into other object-oriented models. This presentation format will be both natural and expressive for domain experts, and easily converted to other machine readable formats, such as RDF (Resource Description Framework) and XML. Considerable effort has gone into achieving these goals; all cross-references and inheritance of properties, for example, are explicitly resolved. This has led to a certain degree of redundancy, but makes the document more readily comprehensible and facilitates its use as a reference document.

The standard is intended to cover all concepts relevant to cultural heritage information, but most particularly, those needed for wide area data exchange between museums, libraries, and archives. Due to the diversity of the subject matter and the nature of cultural heritage collections, this goal can be achieved only by extensions to the standard. However, thanks to its object-oriented nature, the ontology offers a backbone of powerful general concepts, which have a wide application.

The primary purpose of ISO 21127 is to offer a conceptual basis for the mediation of information between cultural heritage organizations such as museums, libraries, and archives. The standard aims to provide a common reference point against which divergent and incompatible sources of information can be compared and, ultimately, harmonized.

The mechanisms involved are the cultural organization and the technical project team. The reason why this cooperation is required is that this activity is both technical—it will shape the technical development of the e-service—but is also highly connected to the conservation practices. Only the experts in the management of the cultural content can, in fact, specify how and what parts of the content would need to be shared.

The output of this activity is the systems specification document, enriched with the information about interoperability.

**Cultural Content Organization**

In this activity, the content is organized. In particular, the cultural organization (the mechanism) takes care of creating the descriptive data about the digitized material: the metadata. The description is based on the specification document and takes into consideration the interoperability needs specified in the previous activities.

The specific activities that we have identified in our survey of existing e-services are the following:

- Primary data registration and description of cultural objects. This data refers to basic identification information, classification, physical shape, condition, geographical location, construction date, application, properties, and relation with other objects.
- Collection of the administrative information concerning the relevant collection and monument management. This information is the prerequisite for the management and documentation of monuments or objects, which are part of a museum collection.
- Documentation of digital products and processes. The meta-data that concern the processes of digitization, identification, quality and thematic content of the digitized material belong to this category.
- Preservation information of the cultural content. This information refers to meta-
data for the long-term preservation of the digitized material.

- Publishing data of the cultural content. The digitized material is published on the Internet and is also stored in optical technology media (CD-ROMs, DVDs). The data included in this category aims at the educational use of the cultural digitized material and the effective creation of multilingual versions.

Considering these activities, and the specification document we recommend, as control, the use of the ICOM/CIDOC standards (Abraham & Means, 2001; Aquarelle, 1999; Baskerville & Pries-Heje, 2004). The ICOM/CIDOC standards are the output of a joint effort across museums, archaeological sites, and other cultural initiatives to create internationally accepted guidelines for museum object information. Since the ICOM/CIDOC standards have been agreed by the cultural organizations, their use should be straightforward.

The output of this activity is the final requirement document containing functional, structural, and process specifications.

**E-Service Development**

Once the requirements have been collected, the material digitized, and the standards set, the time for the actual e-service development arrives. We envision the e-service to be, as a starting point, of the kind presented in the incunabula project: a complete, widely accessible, shared content of precious cultural material.

The e-service development can be regarded as a problem of information system development. In the particular case of cultural heritage e-services, there are particular conditions that make them different in relation to the needs of traditional Web sites treated in the literature (Baskerville and Pries-Heje, 2004). The needs for quick adaptability and flexibility (ibid) do not apply in most cultural heritage cases where the data is stable and the user interfaces are also quite stable in time. The presentation of the artworks is also quite stable in time and it has achieved an agreed upon format. Furthermore, the digitization might involve very fragile artworks, besides being very complex and time consuming, and therefore, it cannot be envisaged that the basic data coming from digitization can be recreated multiple times.

Therefore, the development should not focus on rapid and continuous development, but rather, comply with the more traditional system development lifecycle, focusing on long term usage: efficiency, reliability, low cost maintenance, and robust functionalities (Truex, Baskerville, and Klein, 1999). However, given that most e-services will be Web-based, some of the characteristics of Web sites can be used advantageously in cultural heritage services. The interface must be particularly cured as visitors are varied in age and education; therefore, particular attention to usability must be taken. Furthermore, standards must be strictly respected as presented previously, since data might and should be used in multiple sites or in multiple ways in the same site. Being prepared to provide multiple views of the same digitized material will be a critical factor for the success of cultural heritage e-services over time because the development of views and interfaces will demand much less work compared to the digitization of content. Finally, particular attention should be posed to scalability of e-services, given the staggering quantity of material potentially available for display.

Our case and our survey show, for most e-services, the first user group to be served are professionals. Serving professionals—with high requirements on the quality of the digitization and on the metadata—are a very good starting point because, for the reasons connect to the aforementioned material, it is easier to simplify a high-quality digitization, rather than re-digitize to add information.

Therefore, we envision at least a double loop development where, in a first instance, the
Development Of E-Government Services For Cultural Heritage

cultural organizations and the professionals are served. Once professional e-services are created, then a second loop of development can be carried out where other services are added for the wider public. The first loop uses as mechanism the technical team, the cultural organization, and the professional users. The second loop uses the technical team and representatives of the public as mechanisms in the first activity.

The control proposed for this activity is the ISO 9126 standard, which is concerned primarily with the definition of quality characteristics to be used in the evaluation of software products. Using this standard as control should assure the quality of the e-service.

The output of this activity is one or multiple e-services.

REFLECTION ON THE FRAMEWORK AND ISD METHODOLOGIES

At first sight, the framework resembles a variation of the waterfall methodology for information system development. This impression is undoubtedly strengthened by the fact that the e-services for cultural heritage are indeed information systems. However, the framework proposed in Figure 15 does not propose a waterfall-like development, but rather, shows the activities involved with their inputs, controls, mechanisms, and outputs. The framework does not specify the sequence of the activities, even though visually it might seem so. The fact that the activities are depicted in Figure 15 from the top left to the bottom right follows simply a tradition in IDEF0 modelling.

In fact, in real cases, the activities might go on in parallel or even be carried out in different moments. Typically such an e-service can use material digitized long before the project is started or, even more commonly, can use descriptions (metadata) created before the project is started. The framework (as it does not specify a sequence) can be adopted in any ISD methodology and will still be able to provide indications for what to do, following which rules, who has to do what, and what input has to be used.

The order in which the activities are actually carried out, during the project of e-service development, should be decided contextually by the actors involved. Our research has pointed out the importance of user involvement, where users are not only professionals, but also and foremost, laymen. Appealing to the wider laymen category will, in fact, ensure the long term interest in the e-service. Therefore, the use of a process that contains iterations based on user testing is highly recommended and the framework will guide the choice for inputs, outputs, controls, and mechanisms for each activity in the multiple iterations.

CONCLUSION

The study of the existing initiatives of e-government on cultural heritage shows that the different member states are pursuing the Lisbon strategy for eEurope, but in different ways and at different paces. While some countries are more advanced than others, even for the leading ones, it is difficult to speak about concerted action to create a one-stop site for cultural heritage being this at subject level (archaeology, graphical art, sculptures, and so forth), local level, national level, or international level. While cooperation obviously exists among museums, ministries, universities, and other institutions, this cooperation only very recently has begun to be addressed seriously. While cultural organizations, research institutions, and standard organizations have begun to create usable standards, one of the main issues that we have observed in an initial survey is the lack of a methodology that systematically integrates and adopts these standards.

The in depth examination of three case studies of e-services has pointed out four dimensions that are important to address when developing these
Development Of E-Government Services For Cultural Heritage

services: digitization, requirements engineering, standardization, and interoperability.

These dimensions have been linked to the existing standards to create a development framework, which addresses both practical issues and future needs.

In our view, the problem of advancement in the cataloguing and divulgence of electronic material does not reside in technology, but in organizational traditions and cultures that, knowingly or unknowingly, might be retarding the process, and in the recognition that electronic cataloguing and presentation of artwork is not locally limited (e.g., at museum or site). Politicians and administrators need to understand that to achieve the goals of eEurope, cultural heritage needs to move from the physical place to the electronic space. We believe that this article provides a solid framework to achieve this goal. The proposed framework is easy to understand: based on IDEF0, clear in its content and in the definition of what is needed, who is needed, what rules apply, and what are the goals for each activity. It provides, to the willing and open-minded team, the right guidance (without being restrictive) towards the development of e-services for cultural heritage that are not only effective upon completion for a specific location, but that will stay effective in the future and without borders, if the standards for interoperability are respected.

This framework should be further examined in an action research environment where not only the framework should be put into use but the organizational conditions that characterize this type of projects could be studied, as well. This will allow us not only to validate the framework but also to study the condition under which the “cultural heritage e-service plan” (the primary input of the framework) is actually made.

REFERENCES


Web Sites Referenced (all Web sites visited on November 14, 2004):

Direzione Generale per i Beni Archeologici – Italy: Virtual visit of the Villa Adriana http://www.archeologia.beniculturali.it/pages/atlantearcheo/1_VAD/pano/Atla_Pano_VAD_08.html

Uffizi Gallery, Florence, Italy: http://www.uffizi.firenze.it

The Louvre Museum, Paris, France: http://www.louvre.fr/


Project Piero della Francesca: http://www.piero-dellafrancesca.it


Ministries and Departments of Culture Visited (in alphabetical order: Web sites visited on November 14, 2004):

Denmark—Kulturministeriet: http://www.kulturministeriet.dk/sw16739.asp (press release about the law on conservation of cultural heritage through digitization in Denmark)
Development Of E-Government Services For Cultural Heritage

France—Ministère de la Culture et de la Communication: http://www.culture.fr/ (specific site dedicated to cultural activities)

Greece—Hellenic Ministry of Culture: http://www.culture.gr/

Ireland—The Department of Arts, Sport and Tourism: http://www.arts-sport-tourism.gov.ie/

Italy—Ministero per i Beni e le Attività Culturali: http://www.beniculturali.it/


United Kingdom—Department for Culture, Media, and Sport: http://www.culture.gov.uk/default.htm

References for the Cultural Content Part:


Text Encoding Initiative (TEI) http://www.teic.org/


Sharable Content Object Reference Model (SCORM), http://www.adlnet.org/index.cfm?fuseaction=SCORMDown

Development Of E-Government Services For Cultural Heritage

ENDNOTES

1 English, Italian, Greek, French, Danish, Swedish, German, Spanish

2 Prof. Kiernan, University of Kentucky, is considered the highest authority in the Beowulf manuscript and he was the main figure in the Electronic Beowulf Project. His website—http://www.uky.edu/~kiernan/eBeowulf/guide.htm—provides very extensive material about the project.

3 Information coming from http://www.pierodellafrancesca.it

4 In the Piero della Francesca project, we highlighted the existence of the games. This is an example of customization of the content to a young audience.

Chapter 2.10

An Activity–Oriented Approach to Designing a User Interface for Digital Television

Shang Hwa Hsu
National Chiao Tung University, Taiwan

Ming-Hui Weng
National Chiao Tung University, Taiwan

Cha-Hoang Lee
National Chiao Tung University, Taiwan

ABSTRACT

This chapter proposes an activity-oriented approach to digital television (DTV) user interface design. Our approach addresses DTV usefulness and usability issues and entails two phases. A user activity analysis is conducted in phase one, and activities and their social/cultural context are identified. DTV service functions are then conceived to support user activities and their context. DTV service usefulness can be ensured as a result. The user interface design considers both activity requirements and user requirements such as user’s related product experience, mental model, and preferences in phase two. Consequently, DTV usability is achieved. A DTV user interface concept is thus proposed. The interface design concept contains the following design features: activity-oriented user interface flow, remote control for universal access, shallow menu hierarchy, display management, adaptive information presentation, and context sensitive functions. Usability evaluation results indicate that the user interface is easy to use to all participants.

INTRODUCTION

DTV has several advantages over conventional analogue television: better picture and sound quality, more channels, interactivity, and accessibility.
Many countries plan to switch to DTV within the next 5 to 10 years. To facilitate DTV user adoption, Hsu (2005) conducted a survey exploring factors driving DTV diffusion. Identified factors were: government support, reliable technology, price, DTV service usefulness, and easy-to-use user interface. Service usefulness and easy-to-use user interface were among the most important driving factors determining user adoption.

Identifying useful services and designing an easy-to-use user interface is challenging. First, DTV can be complicated for most users. After switching from analogue to digital, television is no longer just a standalone audio-video device. DTV is becoming central to the digital home by providing information services; controlling home appliances and security devices; supporting communication activities; performing business transactions; storing video images; and so on. Therefore, users’ manipulation of numerous service applications and easily accessing what they want is critical. Secondly, nearly everyone may use DTV. However, different users may use DTV services for their own purposes and have different usage patterns. Moreover, users interacting with the system have different experiences and abilities. Therefore, accommodating diversified user needs and skill levels is also important.

The user-centered design (Norman & Draper, 1986) has been a dominant approach to user interface design in meeting usability requirements. The user-centered design approach essentially focuses on and includes the user in the design process, beginning with user identification and user tasks. User’s target-market characteristics are identified according to user analysis. These characteristics include: demographic characteristics, knowledge and skills, limitations, attitude, and preferences. Based on these characteristics, product requirements can be determined. Task analysis identifies user needs in task performance. These needs include functions and information as well as control, enabling users to perform tasks. Task analysis results are used to specify functional and use requirements. The user interface considers compatibility between user’s information processing limitations and task demands, to ensure usability. Since the user-centered approach recognizes that the design will not be right the first time, it suggests that an iterative design be incorporated into the product development process. By using a product prototype, the design can be refined according to user feedback. The user-centered design approach has been successfully applied to many computer products.

However, the user-centered approach has some drawbacks and limitations. Norman (2005) warned against using it in designing everyday products. He argued that the user-centered approach provides a limited design view. It is suitable for products targeted for a particular market and for specific user-task support, but everyday products are designed for everyone and support a variety of tasks. These tasks are typically coordinated and integrated into higher-level activity units. For everyday product design, Norman proposed an activity-oriented design approach. He asserted that a higher-level activity focus enables designers to take a broader view, yielding an activity-supportive design. In a nutshell, the activity-oriented approach focuses on user activity understanding, and its design fits activity requirements.

In line with Norman’s (2005) notion of activity-centered design, Kuutti (1996) proposed an activity theory application framework to the human-computer interaction design (HCI). According to activity theory (Engeström, 1987), an activity is the way a subject acts toward an object. An activity may vary as the object changes, and the relationship between the subject and object is tool mediated. Tools enable the subject to transform an object into an outcome. Furthermore, an activity is conducted in an environment that has social and cultural context. Two new relationships (subject-community and community-object) were added to the subject-object model. The community is a shareholder group in a particular activity or those who share the same activity objective. Rules and
regulations govern how an activity is carried out
in a community and impose environmental influ-
ences and conditions on an activity. Furthermore,
community members assume different activity
roles and responsibilities. These components
provide a social contextual framework for HCI
design. Therefore, activity theory not only focuses
on user activity analysis, but also considers coop-
 erative work and social relationship interactions
(Kuutti, 1996). That is, activity theory treats
HCI as a complete interaction perspective, not
limited to interaction between an individual and
a computer.

Although several researchers (e.g., Bødker,
1996; Nardi, 1996) have attempted to apply ac-
tivity theory to HCI design, there is no formal
method for putting activity theory into practice.
To resolve this problem, Mwanza (2002) proposed
a systematic design framework called activity-
oriented design method (AODM), based on the
activity triangle model (Engeström, 1987). The
method contains eight steps aimed at successively
collecting information regarding eight activity
components: Subjects, objects, tool, rules, com-
munity, division of labor, transformation process
and outcomes. Mwanza successfully applied
activity theory to two projects by employing this
eight-step model.

DTV is intended for individual use in daily
life activities. A life activity can include many
others (family members, friends, or others) be-
sides the individual, and may include many social
processes. Therefore, DTV can be a social device
shared by family and friends alike. DTV design
should consider social and cultural contexts.
Furthermore, life activities are dynamic. They are
composed of task sets that change as conditions
change. Thus, the design should reflect possible
actions and conditions in which people function
(Norman, 2005).

This chapter proposes an activity-oriented
approach to DTV user interface design. This
approach begins with an understanding of life
activities and their requirements in use context.
Activity theory is used as a framework for life
activity understanding. Life activities are treated
as analysis units. Analysis on use context includes
interaction among community members under-
lying the activity, rules governing the activity,
and physical, as well as sociocultural aspects
of the use environment. From activity analysis,
activity requirements are determined and, in
turn, functional requirements can be specified.
After activity requirements are determined, the
approach moves to identifying user requirements
such as users’ experiences, skills, knowledge,
and preferences. User interface design tries to
optimize compatibility between an activity and
its contextual elements so that it not only meets
activity requirements, but also satisfies individual
as well as social needs. In so doing, DTV useful-
ness and usability can be ensured.

The proposed activity-oriented approach
consists of four steps:

- **Analyze DTV activity requirements.** People engage in an activity in order to
attain an objective. The objective is then
transformed into an outcome that motivates
people to engage in the activity. Accordingly,
the activity can be identified by the motive
for using DTV. Since the identified activities
may be too numerous, related activities are
grouped into activity groups for simplic-
ity.

The intertwined relationships among ac-
tivity components in each activity group
are analyzed. All members and their roles
in the underlying activity are identified.
Moreover, rules governing DTV use and the
social processes underlying the activity are
analyzed. Social and cultural issues related
to DTV use are identified. Activity models
are then constructed to depict relationships
between the eight activity components.

- **Envision services to support each activity
group.** By analyzing activity requirements
and its social context, appropriate applica-

tion services and contents are envisioned. The envisioning process is based on activity models constructed in the previous stage. Conceptual scenarios developed describe how the conceived services support user activity. A DTV system conceptual model is then developed describing system services, their attributes, and the relationships between services. The DTV system can then smoothly meet activity requirements and prove useful to users as a result.

- **Establish a use model.** A use model delineates how users will use DTV services to carry out their activity. Use model development begins with detailed activity analysis. An activity has a hierarchical structure containing a coordinated, integrated set of actions. Actions are targeted toward a specific goal and are achieved through operations. Conditions will control what operations will be taken, and conditions and goals together will determine tool use. Based on the activity-actions-operations hierarchy and activity flow pattern, a use model is then developed to describe the mapping between DTV services and activity flow. This mapping is used as a basis to determine user interface structure and components.

- **Understand user requirements.** To achieve usability goal, user interface design must consider user characteristics. User characteristic exploration includes two parts: (1) understanding users’ previous experiences, and (2) exploring users’ mental models and preferences for interaction methods. Users accumulate previous experience from earlier product interaction. Previous experience contains users’ knowledge, skills, mental models, and affective responses to the related products. Leveraging users’ previous related product experiences can ease users’ new product learning and prevent user frustration. Furthermore, users interact with the system based on their system knowledge that is, their mental models. Designing a user interface compatible with a user’s mental model enhances product usability. Interaction method preference affects users’ product attitude. User interface design must meet these user requirements to enhance usability and increase user acceptance.

**CASE STUDY: A DTV USER INTERFACE DESIGN**

Several qualitative and quantitative methods employed in this study applied the activity-oriented approach. Qualitative methods include ethnographic research, in-depth interviews, and a focus group. These qualitative methods analyze activities and their social/cultural context and conceive DTV services. Quantitative methods were also conducted, assessing activity importance and collecting user characteristic information. A survey was conducted to meet the purpose. The survey comprised of questions regarding user activities; services and contents needs; user experiences with related products; user expectations of user interface style; and user characteristics such as demographic variables and attitudes toward technology adoption. The one thousand questionnaire-sampling scheme—distributed across Taiwan—proportionately represented the Taiwanese population in terms of age, sex, occupation, and geographical location. Of the returned questionnaires, 960 were valid.

**Understanding User Activities and Their Context**

Ethnographic research conducted observed current life activity practices. Twelve lead users (Rogers, 1995) participated in the research. Observations focused on activities engaged in; when and where they were engaged in; how the activity was carried out; who took part in the activity and their roles; and what rules governed this activity.
Observers tried to piece information together using the activity triangle model (Engeström, 1987). In some cases, observation results were insufficient to provide a whole activity system picture. Thus, in-depth interviews conducted collected supplementary information. Through ethnographic research and in-depth interviews, 34 life activities were identified and their activity models obtained.

These activities then served as a candidate activity list in the survey questionnaire. Survey respondents were asked to rate the importance and frequency of activity performance. Exploratory factor analysis conducted explored the interrelationships of 37 activities. A Kaiser-Meyer-Olkin (KMO) test of sampling adequacy obtained a good value of 0.911, the Bartlett’s test of Sphericity was significant (p <0.001), and the matrix determinant was in acceptable range. Factor analysis results determined five activity groups (refer to appendix A): (1) entertainment (9 activities), (2) information (6 activities), (3) education (9 activities), (4) transactions (5 activities), and (5) daily living (5 activities).

Identifying DTV Service Needs to Support each Activity Group

A focus group meeting met to envision potential DTV services and contents to support each activity group. The focus group consisted of two DTV R&D engineers and six lead users with DTV experience in other countries. The meeting began with a briefing on the latest DTV technology and service developments by the two R&D engineers. Next, lead users were given the activity model obtained in the earlier stage and asked to brainstorm potential DTV services that would be useful in supporting each activity group. Thirty-seven candidate DTV services and contents were gathered as a result. After identifying potential services, conceptual scenarios were developed, describing service use in carrying out activities.

These 37 candidate DTV services were then included in the survey questionnaire. Respondents in the survey rated the services/contents in terms of their usefulness in supporting life activities. Useful services for each activity group are shown in Table 1. Some services/contents were considered useful in supporting more than one activity group.

Developing a DTV use Model

The second ethnographic research conducted collected detailed activity analysis information to determine identified service and content functional requirements in each activity group. Five families were recruited for the ethnographic research. These families were lead users of 3C products. The research took place on one weekday and one weekend. Observations were made to learn how life activities are performed in the family context and how each family member

<table>
<thead>
<tr>
<th>Table 1. Useful services for each activity group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entertainment</strong></td>
</tr>
<tr>
<td>Video-on-demand</td>
</tr>
<tr>
<td>Music-on-demand</td>
</tr>
<tr>
<td>Games</td>
</tr>
<tr>
<td>Travel</td>
</tr>
<tr>
<td>Horoscope</td>
</tr>
<tr>
<td>TV shopping</td>
</tr>
<tr>
<td>Ticket Purchase</td>
</tr>
<tr>
<td>News-on-demand</td>
</tr>
<tr>
<td>Voting</td>
</tr>
<tr>
<td>Transportation</td>
</tr>
</tbody>
</table>

520
took part in the activities. Observation results gathered decomposed activities into action and operation levels. Moreover, in-depth interviews requested interviewees to envision how they would use DTV services and contents to support their activities. Use models were then developed based on interview results. These use models served as a basis for determining user interface components and its structure.

Ethnographic research results indicated that each activity group had its own specific use model. For example, entertainment activities are performed in a sequential manner. After users find the right program, they will watch the program for a while before they move onto the next activity. Therefore, a sequential menu structure is appropriate to support this type of activity. On the contrary, an information-searching activity is an iterative process. Users navigate through information space in order to seek for related information. Users may easily get lost and find it difficult to return to their original location. Therefore, a network structure is suitable for this type of activity.

ANALYZING USER REQUIREMENTS

Understanding User Experience

Survey respondents rated past experience with related products in terms of ease of use and frequency of use. The ease-of-use score reflects product usability and the frequency-of-use score indicates activity support usefulness. Combining the two scores yielded a good experience index for each product. Results showed that nine multimedia products, four information products, and two communication products were rated as “good experience products.” Among these products, respondents with high interaction skill rated television sets, TV games, and personal computer as the top three “good experience products,” whereas television sets and TV games were regarded as “good experience products” by respondents with low interaction skill.

Identifying UI Design Features

In-depth interviews conducted identified design features that contribute to good experience. Eight persons participated in the in-depth interview sessions, including two businessmen, two children, two housewives, and two older adults. They were frequent users of “good experience products.” In the interviews, participants were asked to describe their interaction experience with TV sets, TV games, and personal computers and then list design features they considered good.

Good design features for TV sets were: (1) control function visibility—all control functions have dedicated physical keys easily located on the remote control; (2) shallow menu hierarchy—most TV menu options are organized in one or two levels. Too many menu layers confuse and disorient users so that they have difficulty reaching the target option; (3) feedback information visibility—TV provides immediate selection and system response feedback. Selection feedback shows immediate user selection. System response feedback provides the function whether the user activates it or not; and (4) automatic setting and display management provision—TV users do not need to memorize and learn how to go back to the initial display management setting. When the user turns off the TV set, it will automatically return to the initial setting when it is turned on again.

TV games are popular entertainment for children. Good design features indicated by the interviewees include: (1) comprehensible icons—users can easily identify menu option icons in games; (2) unambiguous function grouping—games always use a meaningful group function rule so users can easily locate the function; (3) legible text display—TV games fit information presentation into the display space so children can read the
An Activity-Oriented Approach to Designing a User Interface for Digital Television

instruction; (4) easy-to-use function keys—all function keys are unambiguously mapped to the player’s task; (5) context-sensitive help—when playing a TV game, the player does not need to refer to the manual. Help information for each possible problem is always provided to the player throughout the game play.

PC users pointed out several good design features of a PC: (1) interactivity—PC provides interactive functions and dynamic feedback allowing users to use appropriate commands to achieve their task goal; (2) graphical user interface—PC contains easy-to-understand icons and intuitive metaphor that makes users feel they are directly manipulating the device; (3) direct linkage to related information—the hyperlink mechanism helps users easily and efficiently locate information they are interested in; (4) multi-tasking—users can perform more than one task simultaneously. PC users wished that DTV would also allow users multiple viewing of TV programs and services at the same time.

These good design features provide insight into DTV design. They can also be used as design references to ensure interoperability of networked products.

Exploring Users’ Mental Models and Preferences

Respondents with low interaction skills, such as housewives and older adults, stated their preference for TV interaction style. In fact, even respondents with high interaction skill expected DTV to retain a TV context for entertainment activities, while good PC design features provide support for information-related activities. TV design features all respondents would like to retain on DTV were: (1) TV screen as a default screen, (2) remote control as the main interaction device, (3) shallow menu structure, (4) automatic display settings, and (5) direct access function keys.

The User Interface Design Concept

This study developed a user interface concept based on both activity and user requirements. The concept and the features are described next.

Activity-Oriented User Interface Flow

The use model obtained from ethnographic research served as the design basis for the user interface structure. In this design, services are provided in accordance with activity needs. In addition, services are embedded in the activity so that users need only focus on the activity without contemplating how to apply appropriate services to support the activity. For example, when the user chooses the activity he/she wants to engage in, he/she may get related services in the service menu. Also, when the user uses a service to perform an activity, the system will automatically display related functions. In so doing, users’ cognitive load can be reduced.

Remote Control for Universal Access

Survey data showed that most respondents preferred to use a remote control as the primary interaction device. The remote control design is intended to support users with different interaction skills. Thus, in our design, the remote control is divided into two areas: the primary control area and the advanced control area (Figure 1). The primary control area is similar to the traditional analog TV remote control. It provides basic function keys (e.g., number key, volume up/down, channel up/down, mute, and so on) sufficient for operating multimedia devices. The primary control area also provides eight function keys dedicated to frequently performed activities identified from the activity analysis. The function keys allow users direct access to critical services or programs for those activities. The primary control area is designed mainly for supporting entertainment and
An Activity-Oriented Approach to Designing a User Interface for Digital Television

daily life activities. The advanced control area is a touch pad serving several functions: (1) points and selects menu items, (2) inputs text, (3) displays personalized Electronic Program Guide, and (4) manages displays. It is suitable for information, transaction, and education activities.

**Shallow Menu Hierarchy**

To provide a simple use environment, the menu function items were arranged based on importance and frequency-of-use for each service. The sequential menus were limited to three layers, helping users avoid unnecessary menu traversal. For simpler operation, menu items can be selected using arrow keys to move the cursor and then pressing the enter (ok) key. Moreover, eight hot keys on the remote control provide direct access to critical activity group services (Figure 2) so that menu traversal is minimized.

**Multiple Viewing and Display Management**

To accommodate single and multiple viewing behaviors, two display modes (single viewport and multiple viewport) are provided. The single-view mode is designated as a default mode. When the DTV is turned on, it displays single-view mode (Figure 3a). In the single-view mode, if another service or program is selected, the DTV will automatically switch to multiple-view mode (Figure 3b). In the design, the system supports up to four active viewports at the same time. With multiple-view mode display, the user can resize and relocate the viewports at will, or go back to the initial single-view mode by pressing the cancel key on the remote control.

**Adaptive Information Presentation**

In the multiple-view mode, display information legibility may suffer. To solve this problem, our study developed an adaptive information presentation method. The method contains two mechanisms: menu legibility adjustment and content legibility. Menu legibility affects how many menu options can be displayed. The adaptive information presentation method automatically adjusts the numbers of legible options for viewport size to ensure menu option legibility. Content legibility affects how much information can be displayed. The adaptive information presentation...
method decides which information is displayed according to its importance. When a display space is reduced, only important information will be displayed, and font size is automatically adjusted to keep information legible (Figure 3b and 4). In addition to the adaptive information presentation method, the system also provides a virtual magnifying glass that enlarges focused content while it shrinks the neighborhood area using a fisheye display method.

**Context-Sensitive Menu Functions**

To avoid unnecessary menu traversals, functions appropriate to the current use context are provided to users. The context-sensitive menu functions can detect a user’s location in the activity menu context and provide suitable activity support options. The functions are displayed in a menu bar and numbered in terms of their appropriateness (Figure 4).

**USABILITY EVALUATION**

The purpose of usability evaluation is twofold: (1) assessing user interface design usability, and (2) evaluating usefulness of each user interface design feature from the user’s point of view.

**Participants**

Twenty subjects participated in the usability evaluation. In order to represent user population,
the recruited participants included 11 high skill level subjects (i.e., three high school students, four college students, four middle-aged technical professionals) and nine low skill level subjects (i.e., three children in elementary school, three middle-aged housewives, and three older adults). They all regarded themselves as early adopters of DTV when they answered the survey question.

**Material, User Tasks and Procedures**

A user interface concept prototype was developed using Microsoft Visual C++. The prototype simulated the user interface to support three kinds of activities: entertainment, transaction, and information browsing. A liquid-crystal display (LCD) TV was used to display the prototype. The experimental environment is shown in Figure 5.

Before the usability evaluation session began, the experimenter introduced the evaluation procedure and tasks to the participants. In the evaluation stage, subjects were asked to perform tasks by following a test scenario. The test scenario represented activities frequently engaged in by family members in a morning. It consisted of three scenarios: (1) searching for news/movie program, (2) browsing information, and (3) shopping on TV. Each activity scenario in turn contained several tasks (Table 2). Specifically, the searching for news/movie program scenario involved three tasks: (1) selecting a TV program from the program guide, (2) browsing a movie from a video-on-demand (VoD) service, and (3) scheduling a program recording. The information-browsing scenario illustrated daily information-related activities such as checking weather, stock information, and traffic information. Finally, the TV shopping task simulated TV-shopping behavior including searching for products, browsing product information, comparing different products, and engaging in online transactions.

When performing these tasks, subjects were asked to think aloud through every task step. After performing the tasks, performance data (task error rate and verbal protocol) were recorded. Upon each task completion, subjects rated the perceived usefulness (five-point Likert scale: 1-non-useful, 5-very useful) and perceived usability (five-point Likert scale: 1-unusable, 5-easy to use) of each user interface feature.

**RESULTS**

**Task Performance**

**Scenario 1: Searching for News/Movie Programs**

In the searching for news/movie programs scenario, almost all operations were error free, except that a few subjects made an error at the

---

Figure 5. The experimental environment
An Activity-Oriented Approach to Designing a User Interface for Digital Television

step “press function key (error rate = 0.4)” in the scheduling a program recording task (Table 3). In order to understand why subjects made such an error, we had a debriefing interview with subjects who made the error. We found that the low skill group subjects expected to find the “record” key on the remote control, while the high skill group subjects anticipated that the “record” function was in the main menu on the screen. These expectations were not consistent with our design. In our design, the “record” function is only embedded in the context sensitive function bar. The result caused us to reconsider adding an array of the hard function keys on the remote control for those frequently used functions (e.g., record, play, stop, forward, back).

Scenario 2: Browsing Information

There is only one operational error found in this scenario (Table 4). Subjects made errors in the “select the information item” step of the information-browsing task (error rate = 0.15). After completing the task, we interviewed the subjects and found that three low-skilled subjects did not understand what “submenu” means, because they

Table 2. User’s activities, tasks, and operations

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Sub Tasks</th>
<th>Operations &amp; [DTV design features in the task]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searching for a News/Movie program</td>
<td>selecting a TV program</td>
<td>(1) power on; (2) key in target program number; (3) press enter</td>
</tr>
<tr>
<td></td>
<td>browsing a movie from a video-on-demand</td>
<td>(1) power on; (2) main menu [shallow menu structure]; (3) select VOD; (4) select program type; (5) select program; (6) start movie</td>
</tr>
<tr>
<td></td>
<td>scheduling a program recording</td>
<td>(1) power on; (2) key in target program number; (3) press enter; (4) press function key [context sensitive functions]; (5) select function; (6) start recording</td>
</tr>
<tr>
<td>Browsing Information</td>
<td>checking weather</td>
<td>(1) power on; (2) main menu [shallow menu structure]; (3) select information services; (4) select the item ‘weathers’ (In TV viewing state, use multiple display): (1) press short-cut bar; (2) select items – ‘weather information’ adaptive information presentation]; multiple viewing &amp; display management]</td>
</tr>
<tr>
<td></td>
<td>stock information</td>
<td>(1) power on; (2) main menu [shallow menu structure], (3) select information services, (4) select the item ‘stock’; (5) select a target stock. (In TV viewing state, use multiple display): (1) press short cut bar; (2) select items ‘my stock’ [adaptive information presentation], multiple viewing &amp; display management]</td>
</tr>
<tr>
<td></td>
<td>traffic information</td>
<td>(1) Power on, main menu [shallow menu structure]; (2) select information services, (3) select the item ‘traffic’ (in TV viewing state, use multiple display): (1) press short cut bar; (2) select the item ‘traffic information’; adaptive information presentation], multiple viewing &amp; display management]</td>
</tr>
<tr>
<td>Shopping on TV</td>
<td>TV - shopping</td>
<td>(1) power on; (2) main menu; (3) select purchase; (4) TV-shopping; (5) select shopping store; (6) select target products; (7) compare products; (8) buy the product. [shallow menu structure], [activity oriented UI flow]</td>
</tr>
</tbody>
</table>
did not have any computer experience. In revising the menu design, we considered providing a clue to guide the low-skill user group when they interact with the TV menu. Or, we would also provide a number key input for the low skill group to allow them direct function access.

Scenario 3: Shopping on TV

All subjects not only understood how to perform the TV shopping task but were also able to complete it within a short time. In the debriefing interview, we asked subjects why they were so skilled. Subjects indicated that they felt the TV shopping functions were similar to real life shopping situations. In addition, the function arrangement in the menu structure was easy to understand and navigate.

Evaluation of Usefulness for Each User Interface Design Feature Activity-Oriented User Interface Flow

All subjects were able to follow the user interface structure to perform TV viewing, information retrieving, and transaction activities with minimal errors (mean error rate = 0.03%). Subjects felt that the activity-oriented user interface flow was useful (mean of perceived usefulness = 4.50, Std. = 0.83) and usable (perceived ease of use = 4.30, Std. = 0.98) for the TV-shopping task. With this design feature, they did not have to search for the service they needed at every task step. That is, the system automatically provided relative information and functions to support the TV-shopping task without traversing through multiple service menus. Therefore, they were not afraid that they might get lost while performing the task.

Table 3. Error rate of select news/movie programs tasks

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Sub Tasks</th>
<th>Operations</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searching for News/Movie programs</td>
<td>selecting a TV program</td>
<td>(1) power on</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) key in target program number</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3) press enter</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>browsing a movie from video-on-demand</td>
<td>(1) main menu</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) select ‘Video On Demand’</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3) select program type</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4) select program</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6) start movie</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>scheduling a program recording</td>
<td>(1) key in target program number</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) press enter</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3) press function key</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4) select function</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5) start recording</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 4. Error rate of browsing information task

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Sub Tasks</th>
<th>Operations</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>browsing information</td>
<td>Access services</td>
<td>(1) power on</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) main menu [shallow menu structure]</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3) select information services</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4) select the information item</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(In TV viewing state, use multiple display):</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1) press short-cut bar</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) select services items</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Shallow Menu Structure

Twenty users gave a high rating of perceived usefulness (mean = 3.60, Std. = 1.23) and perceived usability (mean = 3.80, Std. = 1.11) to this design feature when they completed three information retrieving, VoD, and TV-shopping activities. Subjects were able to complete tasks without making any errors (mean error rate = 0.00). Time recordings of menu navigation indicated that the subjects did not need much time thinking about where to go next. Furthermore, the verbal protocol did not show that subjects experienced any frustration or confusion.

Multiple Viewing and Viewport Management

The multiple-viewing feature was evaluated by the information browsing activity. Technical professionals and students exhibited higher acceptance (i.e., mean of perceived usefulness = 4.07 Std. = 0.37; mean of perceived ease of use = 4.36, Std. = 0.74) of the multiple-viewing mode, while housewives (mean of perceived usefulness = 2.33, Std. = 0.58) and older adults (mean of perceived usefulness = 1.67, Std. = 0.58) had problems with this feature. The former expressed that the multiple-view mode was convenient for them because they could use the information service while they were watching a TV program. However, the latter just wanted to use single-view mode. If multiple-view mode were the only choice, housewives and older adults hoped that the system manufacturer could provide an automatic display management function. Although this design feature may not be well received by housewives and older adults, their task performance was still above the baseline (mean of perceived ease of use = 3.50, Std. = 0.58).

Adaptive Information Presentation

This feature was well received by all subjects (mean of perceived usefulness = 3.85, Std. = 1.27) in information-browsing tasks, especially for older adults. It can enhance displayed information legibility. However, a concern was raised that adaptive information presentation might cause a side effect, that is, increasing font size may decrease the number of menu options displayed and consequently hamper menu interaction performance. However, usability evaluation results indicated that this was not the case. Results revealed that it did not have a negative effect on menu selection or searching for menu options (mean of perceived ease of use = 3.85 Std. = 1.27). In fact, it enhanced menu interaction performance for older adults.

Context-Sensitive Functions

All subjects gave high marks to this design feature. Students and housewives expressed that this design feature was helpful (mean of perceived usefulness = 3.80, Std. = 1.01) and usable (mean of perceived easy to use = 4.00, Std. = 1.08) be-

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Sub Tasks</th>
<th>Operations</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shopping on TV</td>
<td>TV – shopping</td>
<td>(1) power on</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) main menu</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3) select purchase</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4) TV-shopping</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5) select shopping store</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6) select target products</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7) buy the product</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 5. Error rate of TV shopping task
cause it could alleviate efforts to traverse menus and reduce operation time.

CONCLUSION

This study attempts to apply an activity-oriented design approach to digital TV. This approach analyzes activities and identifies requirements for each type of activity. Activity-oriented user interface flow can help users complete their activities and reduce their workload. The evaluation results support this notion, and services are considered to be accessible to users and useful in supporting activities.

This study also explores good design features of related products used by users for the same activities. Incorporating good design features of related products into the new design not only eases the learning process for the new product, but also helps establish a familiar and comfortable feeling for first-time users. This helps users gain out-of-the-box experience (Ketola, 2005).

We also found that different user groups have different user interface expectations. For example, technicians and students prefer multiple viewing and multiple tasks. On the contrary, older adults prefer a simple TV environment. Therefore, the proposed design concept is flexible enough to accommodate these two types of interaction methods in order to meet two separate user requirements.

ACKNOWLEDGMENT

This research is supported by MediaTek Research Center, National Chiao Tung University, Taiwan.

REFERENCES


An Activity-Oriented Approach to Designing a User Interface for Digital Television


### APPENDIX: FACTOR LOADING OF EACH ACTIVITY GROUP

<table>
<thead>
<tr>
<th>Activity Group (5 factors)</th>
<th>Information</th>
<th>Education</th>
<th>Entertainment</th>
<th>Daily living</th>
<th>Transaction</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>.788</td>
<td>.191</td>
<td>.173</td>
<td>.243</td>
<td>.413</td>
<td>Ball gaming</td>
</tr>
<tr>
<td>Watching movie</td>
<td>.704</td>
<td>.454</td>
<td>.385</td>
<td>.064</td>
<td>.488</td>
<td>Body-building</td>
</tr>
<tr>
<td>Newspaper</td>
<td>.673</td>
<td>.200</td>
<td>.270</td>
<td>.309</td>
<td>.335</td>
<td>Dinner party</td>
</tr>
<tr>
<td>Bookstore shopping</td>
<td>.616</td>
<td>.257</td>
<td>.203</td>
<td>.193</td>
<td>.268</td>
<td>Camping</td>
</tr>
<tr>
<td>Shopping</td>
<td>.502</td>
<td>.175</td>
<td>.289</td>
<td>.142</td>
<td>.029</td>
<td>Ball game watching</td>
</tr>
<tr>
<td>Magazine</td>
<td>.440</td>
<td>.228</td>
<td>.301</td>
<td>-.147</td>
<td>.368</td>
<td>Singing (Karaoke)</td>
</tr>
<tr>
<td>Information showing</td>
<td>.372</td>
<td>.665</td>
<td>.500</td>
<td>.015</td>
<td>.475</td>
<td>Religious activity</td>
</tr>
<tr>
<td>Social service</td>
<td>.127</td>
<td>.659</td>
<td>.451</td>
<td>.062</td>
<td>.310</td>
<td>Housekeeping</td>
</tr>
<tr>
<td>Art exhibition</td>
<td>.470</td>
<td>.611</td>
<td>.490</td>
<td>-.007</td>
<td>.483</td>
<td>Computer gaming</td>
</tr>
<tr>
<td>Accomplishment course</td>
<td>.294</td>
<td>.594</td>
<td>.242</td>
<td>.262</td>
<td>.398</td>
<td>Pet breeding</td>
</tr>
<tr>
<td>School</td>
<td>.239</td>
<td>.587</td>
<td>.370</td>
<td>.267</td>
<td>.252</td>
<td>Gardening</td>
</tr>
<tr>
<td>Refresher course</td>
<td>.312</td>
<td>.565</td>
<td>.294</td>
<td>.153</td>
<td>.378</td>
<td>Cooking</td>
</tr>
<tr>
<td>Photography</td>
<td>.395</td>
<td>.547</td>
<td>.489</td>
<td>.210</td>
<td>.414</td>
<td>Commodity buying</td>
</tr>
<tr>
<td>Tai-Chi chuan</td>
<td>-.108</td>
<td>.471</td>
<td>.250</td>
<td>.178</td>
<td>.187</td>
<td>TV / e-shopping</td>
</tr>
<tr>
<td>Handicraft making</td>
<td>.214</td>
<td>.445</td>
<td>.361</td>
<td>.263</td>
<td>.106</td>
<td>Investment</td>
</tr>
<tr>
<td>Traveling</td>
<td>.422</td>
<td>.462</td>
<td>.611</td>
<td>.063</td>
<td>.397</td>
<td>Finical news</td>
</tr>
<tr>
<td>Joy riding</td>
<td>.413</td>
<td>.311</td>
<td>.559</td>
<td>.325</td>
<td>.341</td>
<td>Lottery</td>
</tr>
</tbody>
</table>

531
ABSTRACT

The introduction of computing and communications technologies within cars raises a range of novel human-computer interaction (HCI) issues. In particular, it is critical to understand how user-interfaces within cars can best be designed to account for the severe physical, perceptual and cognitive constraints placed on users by the driving context. This chapter introduces the driving situation and explains the range of computing systems being introduced within cars and their associated user-interfaces. The overall human-focused factors that designers must consider for this technology are raised. Furthermore, the range of methods (e.g., use of simulators, instrumented vehicles) available to designers of in-car user-interfaces are compared and contrasted. Specific guidance for one key system, vehicle navigation, is provided in a case study discussion. To conclude, overall trends in the development of in-car user-interfaces are discussed and the research challenges are raised.

INTRODUCTION

The motor car is an integral part of modern society. These self-propelled driver-guided vehicles transport millions of people every day for a multitude of different purposes, for example as part of work, for visiting friends and family, or for leisure activities. Likewise, computers are essential to many peoples’ regular lives. It is only relatively recently that these two products have begun to merge, as computing-related technology is increasingly implemented within road-going vehicles. The functions of an in-car computing system can be broad, supporting tasks as diverse as navigation, lane keeping, collision avoidance, and parking. Ultimately, by implementing such systems car manufacturers aim to improve the safety, efficiency, and comfort and entertainment of the driving experience (Bishop, 2005).

Designing the user-interface for in-car computing systems raises many novel challenges, quite unlike those traditionally associated with interface design. For instance, in many situations, the use of an in-car system is secondary to the complex and already demanding primary task of safely
controlling a vehicle in 2D space, whilst simultaneously maintaining an awareness of hazards, largely using the visual sense. Consequently, the level of workload (physical, visual, and mental) when using displays and controls becomes a critical safety-related factor. As a further example, in-car computing systems have to be used by a driver (and possible also, a passenger) who is sat in a constrained posture and is unlikely to be able to undertake a two handed operation. Therefore, the design (location, type, size, etc.) of input devices has to be carefully considered, accounting in particular for comfort, as well as safety, requirements.

This chapter aims primarily to provide the reader with an overall awareness of novel in-car computing systems and the key HCI design and evaluation issues. The focus is on the user-interface, that is, “the means by which the system reveals itself to the users and behaves in relation to the users’ needs“ (Hackos & Redish, 1998, p.5). Topics of relevance to both researchers and practitioners are raised throughout. Given the complexity of the driving task and the wide range of computing systems of relevance, the chapter principally provides breadth in its consideration of the subject. Nevertheless, some depth is explored in a case study investigation on the design and evaluation of user-interfaces for vehicle navigation systems.

**TYPES OF IN-CAR COMPUTING SYSTEMS**

Technology is increasingly being seen to have a critical role to play in alleviating the negative aspects of road transport, such as congestion, pollution and road traffic accidents (Bishop, 2005). Many technological initiatives are considered under the umbrella term, intelligent transport systems (ITS), where “ITS provides the intelligent link between travelers, vehicles, and infrastructure“ (www.itsa.org, September, 2006). In this respect, in-vehicle computing systems are an important facet of ITS. Specifically, there are two core types of computing and communications systems which are either being implemented or developed for use in vehicles:

- **Information-based systems:** These systems provide information relevant to components of the driving environment, the vehicle or the driver. Examples of systems include navigation (facilitating route planning and following), travel and traffic information (traffic conditions, car parking availability, etc.), vision enhancement (providing an enhanced view of the road ahead, when driving at night, in fog or in heavy rain), driver alertness monitoring (informing the incapacitated driver if they are unfit to drive) and collision warnings (presenting warnings or advice regarding hazards).

- **Control-based systems:** These systems affect the routine, operational elements of the driving task. Examples of systems include adaptive cruise control (where the car is kept at a set time gap from a lead vehicle), speed limiting (the car speed cannot exceed the current limit), lane keeping (the driver’s vehicle is kept within a given lane), self parking (vehicle automatically steers in low speed operation to position itself within a selected parking space) and collision avoidance (the vehicle automatically responds to an emergency situation). Clearly, such systems fundamentally change the nature of what we consider to be ‘driving.’

It is important to note that there is a third category of in-car computing system, those which do not provide any functionality to support the driving task. These systems are an important consideration though, as they can negatively influence safety, particularly through the potential for distraction (Young, Regan & Hammer, 2003). Such systems may aim to enhance work-oriented productivity.
whilst driving (e.g., mobile phones, e-mail and Internet access) or be primarily conceived for entertainment and comfort purposes (e.g., music and DVD players, games). Moreover, they may be designed for dedicated use in a vehicle or for operation in a range of different contexts (often termed nomadic devices).

OVERALL HUMAN FACTORS ISSUES

Driving is a complex task involving a large number of subtasks that can be conceptualised as existing within three levels of an overall hierarchical structure (Michon, 1985):

- Strategic tasks (highest level global travel decisions—e.g., which car to take, which route to take);
- Tactical tasks (making concrete manoeuvres requiring interaction with other road users—e.g., changing lane, turning at a roundabout);
- Operational tasks (motor execution of tasks planned at higher levels—e.g., turning steering wheel, pressing brake).

Inevitably, the introduction of new technologies into the driving context will have a considerable impact across all three levels. As a result, there are many human-focused issues that must be considered in the design and evaluation process for in-car computing systems. To provide structure to a discussion of these issues, two overall scenarios are envisaged which may arise from poor design and/or implementation of the technology.

- **Overload:** Many of these systems (particularly those providing novel types of information and/or interactions) lead to situations in which a driver must divide their attention between core driving tasks (e.g., watching out for hazards) and secondary system tasks (e.g., inputting information). Furthermore, systems may provide excessive information in an inappropriate way leading to high levels of mental workload, stress and frustration. Such issues often manifest themselves as distraction to the driver (biomechanical, visual, auditory and/or cognitive).

- **Underload:** Control-based systems clearly automate certain aspects of driving, transferring certain responsibilities from operator to computer (e.g., staying in lane), whilst potentially providing new tasks for the driver (e.g., monitoring system performance). Automation is a fundamental human factors topic with a considerable research literature (see Wickens et al., 2004). Key concerns in this context relate to the potential for a driver exhibiting reduced situational awareness (e.g., for other road users), negative behavioral adaptation (e.g., by taking greater risks) and de-skilling (e.g., driver not able to resume control in the event of system failure).

THE HUMAN-CENTRED DESIGN PROCESS

The fundamental components of a human-focused approach hold true for in-car computing, as much as for any interactive product or system, that is, early focus on users and tasks, empirical measurement and iterative design (Gould & Lewis, 1985). A comprehensive understanding of the context in which in-car computing devices will be used is especially important early in the design process. Context of use refers to “the users, tasks and equipment (hardware, software, and materials), and the physical and social environments in which a product is used” (Maguire, 2001, p.457).

A context of use analysis assists in developing the initial requirements for a design and also provides an early basis for testing scenarios. Moreover, context of use analysis provides a focused ap-
Designing and Evaluating In-Car User-Interfaces

proach that helps to ensure a shared view among a design team. In the driving situation, there are several context of use issues which will have a significant effect on how an in-car computing system is subsequently designed. Accounting for these raises many unique challenges for in-car user-interface designers.

Users

As with many other consumer products, there will be a large variability in user characteristics (e.g., in perceptual and cognitive abilities, computer experience, anthropometry) to consider when designing in-car computing systems. Car manufacturers may have particular socio-economic groups in mind when designing a vehicle, but the user base may still be extremely large.

One fundamental individual difference factor often addressed in research is driver age—drivers can be as young as 16 (in certain countries) and as old as 90. In this respect, younger drivers may be particularly skilled in the use of computing technology, in comparison with the population at large, but are especially prone to risk taking (Green, 2003). Moreover, studies have shown a limited ability to divide attention and prioritize sources of information, largely due to lack of driving experience (Wickman, Nieminem & Summala, 1998). Subsequently, system block outs, which prevent the use of complex functions in inappropriate driving situations, are likely to be of particular benefit for these individuals.

In contrast, older drivers often suffer from a range of visual impairments that can lead to a range of problems with in-vehicle displays. For instance, presbyopia (loss of elasticity in the lens of the eye) is extremely common amongst older people, as is reduced contrast sensitivity. Studies consistently show that older drivers can take 1.5 to 2 times longer to read information from an in-vehicle display compared to younger drivers (Green, 2003). Given that drivers have a limited ability to change the distance between themselves and an in-vehicle display, the size, luminance and contrast of presented information are obviously critical design factors.

Tasks

A key task-related issue is that the use of an in-car computing system is likely to be discretionary. Drivers do not necessarily have to use the system to achieve their goals and alternatives will be available (e.g., a paper map, using the brake themselves). As a result, the perceived utility of the device is critical. Furthermore, drivers’ affective requirements may be particularly important. In certain cases, this requirement may conflict with safety-related needs, for instance, for a simple, rather than flashy or overly engaging user-interface.

The factor that most differentiates the driving context from traditional user-interface design is the multiple-task nature of system use, and in this respect, there are two critical issues that designers must take into consideration. The first concerns the relationship between primary driving tasks and secondary system tasks, as drivers seek to divide their attention between competing sources of information. Driving is largely a performance and time-critical visual-manual task with significant spatial components (e.g., estimating distances). Consequently, secondary tasks must not be overly time-consuming to achieve or require attentional resources that are largely visual, manual, and spatial in nature, if they are to avoid having a significant impact on primary driving.

A second fundamental issue is the amount of information processing or decision making required for successful task performance, known as mental workload (Wickens et al., 2004). Novel in-car computing systems may provide functionality of utility to a driver or passengers, but interaction with the technology will inevitably increase (or in some cases decrease) overall workload. Context is very important here, as driving is a task in which workload varies considerably
Designing and Evaluating In-Car User-Interfaces

from one situation to another (compare driving in city traffic versus on the motorway). In this respect, certain authors (e.g., Green, 2004; Jones, 2002; Markkula, Kutila, & Engström, 2005) have taken the view that workload managers must be developed which make real-time predictions of the workload a driver is under and only present information or enable interactions to occur when overall workload is considered to be at an acceptable level. As an example, an incoming phone call may be sent straight to voice mail when the driver is considered to be particularly loaded (e.g., when driving in an unfamiliar city), but may be permitted in a lower workload scenario (e.g., driving along a dual carriageway and following a lead vehicle). Simple workload managers already exist in some vehicles (e.g., http://driving.timesonline.co.uk/article/0,,12929-2319048,00.html, September 2006), nevertheless, there are several complex research issues which must be addressed to fully realize the benefits of adaptive software in this context. For instance, workload managers need a comprehensive and accurate model of the driver, driving tasks and the driving environment. Given the vast range of variables of relevance to these categories, many of which do not lend to accurate and reliable measurement, extensive workload managers are likely to remain in the research domain for several years.

Equipment

The driving situation necessitates the use of input and output devices which are familiar to the majority of user-interface designers (pushbuttons, rockers, rotaries, LCDs, touchscreens, digitized or synthesized speech), together with equipment which is perhaps less known. For instance, there is a considerable research literature regarding the use of Head-Up Displays (HUDs) within vehicles. A HUD uses projection technology to provide virtual images which can be seen in the driver’s line of sight through the front windscreen (see Figure 1). They are widely used within the aviation and military fields, and are now beginning to be implemented on a large-scale within road-based vehicles. HUDs will potentially allow drivers to continue attending to the road ahead whilst taking secondary information more quickly (Ward & Parkes, 1994). As a consequence, they may be most applicable to situations in which the visual modality is highly loaded (e.g., urban driving), and for older drivers who experience difficulties in rapidly changing accommodation between near and far objects (Burns, 1999).

From a human-focused perspective, there are clear dangers in simply translating a technology from one context to another, given that vehicle-based HUDs will be used by people of varying perceptual and cognitive capabilities within an
Designing and Evaluating In-Car User-Interfaces

environment where there is a complex, continually changing visual scene. Specifically, researchers have established that poorly designed HUDs can mask critical road information, disrupt distance perception and visual scanning patterns, and negatively affect the ability of drivers to detect hazards in their peripheral vision (known as perceptual tunneling)—summarized by Tufano (1997) and Ward and Parkes (1994). Critical design factors that emerge from these findings include: display complexity; contrast and luminance; color choice; size of image; spatial location; and virtual image distance. Perhaps the most important design-related requirements are to consider carefully what and how much information is most appropriate to present on a HUD. There are temptations for designers to present ever-increasing amounts of information on HUDs. However, in contrast with traditional in-vehicle displays, a HUD image, by its very presence in the driver’s line of sight, will demand focused attention (Burnett, 2003).

Environments

The physical environment is also a specific area that designers need to be aware of. In particular, the light, sound, thermal and vibration environment within a car can be highly variable. A range of design requirements will emerge from a consideration of these factors, for instance, potential for glare, problems with speech interfaces, use with gloves, and so on.

From anthropometric and biomechanical perspectives, the vehicle cabin environment provides many challenges for designers. This is an area in which designers make considerable use of CAD modeling to analyze different locations for displays and controls, ultimately aiming to ensure good fit for the design population. However, drivers sit in a constrained posture, often for several hours and have limited physical mobility (e.g., to comfortably view displays or reach controls). Consequently, there is limited space within a vehicle for the placement of a physical user-interface, a key problem for designers hoping to implement additional functionality within the vehicle.

To a large extent, this factor has fueled the development of multi-modal user-interfaces, where a small number of controls, together with menu-driven screens, provide access to many functions within the vehicle. Clearly, such visually-oriented user-interfaces are likely to promote a considerable amount of “eyes-off-road” time, and empirical studies have confirmed this prediction (Dewar, 2002). Moreover, users mistaking the current mode is a well-established problem in user-interface design, and clear feedback is an important design requirement (Preece, Rogers & Sharp, 2002). In many respects, there is a trade-off in design between the number of discrete controls that a user must scan within a vehicle and the number of levels within a menu-based system that must be explored and understood. This is a very similar problem to that considered by HCI researchers in the 1980s and 1990s interested in the breadth versus depth of menus in graphical user-interfaces (Shneiderman, 1998). An overall recommendation from such HCI research is that breadth should generally be favored over depth, as excessive depth can cause considerably more problems for the user than an equivalent breadth, largely due to the cognitive problems of navigation (Shneiderman, 1998). Whilst such guidance is considered to be of relevance to the design of in-car computing, research is still required which considers the trade-off existing in the multiple-task driving environment.

METHODS FOR USE IN DESIGN AND EVALUATION

In considering the range of methods that a designer can utilize when designing and evaluating in-car computing systems, the first difficulty is in establishing what is meant by a method. In this respect, a “human factors method for testing
Designing and Evaluating In-Car User-Interfaces

in-car systems” can be seen to be a combination of three factors:

1. Which environment is the method used in (road, test track, simulator, laboratory, etc.). As can be seen in Figure 2 (redrawn and adapted from Parkes, 1991), there is a fundamental trade off in choosing a method environment between the need for control and the validity of results. Choosing an environment will also be largely influenced by practical considerations, the knowledge/skills of the design and evaluation team and resource limitations.

2. Which task manipulations occur (multiple task, single task loading, no tasks given, etc.)? In certain methods, there is an attempt to replicate or simulate the multiple task nature of driving. For other methods, performance and/or behavior on a single task may be assessed and the potential impact on other tasks inferred from this. Most removed from actual driving, some methods do not involve users, but instead aim to predict impacts or issues, for instance through the use of expert ratings or modeling techniques.

3. Which dependent variables (operationalized as metrics) are of interest. In assessing an in-car computing user-interface, a large range of possible metrics could be implemented. Some will relate to drivers’ performance with primary driving tasks (e.g., lane position, hazard detection) or their use of primary vehicle controls (e.g., use of brake, steering wheel). Other metrics focus on driver performance and/or the demand of secondary tasks (e.g., task times, errors, display glances). As noted by Parkes (1991), usability evaluations of in-car computing devices should incorporate a wide range of measures relevant to the different levels of the driving task. For instance, at the strategic level, observation techniques and surveys are of relevance, whereas verbal protocols, interviews and questionnaires can capture the behavior of drivers at the tactical level. As noted by Parkes, such an approach provides “complete, rather than partial, pictures of product usability” (p.1445).

There is presently considerable research investigating specific methods for use in the design and evaluation of in-car user-interfaces. As noted by Preece et al. (2002), in deciding on any HCI method, the design team must consider the overall goals of the work, specific questions to be addressed, the practical and ethical issues and how data will need to be analyzed and reported. For

Figure 2. Environments for evaluation of in-car computing devices and the relationship between validity and control
in-car computing, these principles still hold, and many of the same global techniques used in the HCI area (for example, questionnaires, interviews, guidelines/checklists) will be used. However, by necessity, bespoke methods (or at least specific versions of generic methods) are required that account for the particular complex, safety-critical characteristics of the driving context. The following section summarizes key methods currently used and highlights some of the important research issues under investigation. Moreover, primary advantages and disadvantages are given. Table 1 summarizes some of the key issues.

### Field Trials

Participants are given a car fitted with an operational system for several months for use in everyday activities. This method tends to look at broad issues relating to the long-term use of a system, for example, drivers’ acceptance of the technology, and whether any behavioral adaptation effects arise. Objective data can be measured using on-board instrumentation (e.g., cameras, speed sensors) whereas subjective data is often captured using survey or interview-based approaches. Clearly, such a method provides an ecologically valid test of a system, and is particularly

---

**Table 1. Overview of methods used to evaluate the user-interface for in-car computing systems**

<table>
<thead>
<tr>
<th>Method</th>
<th>Environment</th>
<th>Task manipulations</th>
<th>Overall Measures</th>
<th>Primary Advantages</th>
<th>Primary Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field trials</td>
<td>Real road (in everyday driving)</td>
<td>Multi-task (according to driver motivation)</td>
<td>Primary/secondary task performance/behavior, user opinions, etc.</td>
<td>Ecological validity, can assess behavioral adaptation</td>
<td>Resource intensive, ethical/liability issues to consider</td>
</tr>
<tr>
<td>Road trials</td>
<td>Real road (in pre-defined settings)</td>
<td>Multi-task (commonly, evaluator-manipulated)</td>
<td>Primary/secondary task performance/behavior, user opinions, etc.</td>
<td>Balance of ecological validity with control</td>
<td>Resource intensive, ethical/liability issues to consider</td>
</tr>
<tr>
<td>Simulator trials</td>
<td>Virtual driving environment (varying in fidelity)</td>
<td>Multi-task (commonly, evaluator-manipulated)</td>
<td>Primary/secondary task performance/behavior, user opinions, etc.</td>
<td>Control over variables, safe environment, cost-effective</td>
<td>Validity of driver behavior, simulator sickness</td>
</tr>
<tr>
<td>Occlusion</td>
<td>Laboratory/ statically in car</td>
<td>Secondary task achieved in controlled visual experience</td>
<td>Visual demand of user-interface</td>
<td>Standardized approach, control over variables</td>
<td>Limited scope, concern over validity of approach and metrics</td>
</tr>
<tr>
<td>Peripheral detection</td>
<td>Road/virtual driving environment</td>
<td>Multi-task (although commonly, evaluator-manipulated)</td>
<td>Visual/ cognitive workload</td>
<td>Assesses cognitive, as well as visual demand</td>
<td>Can be resource intensive, range of approaches</td>
</tr>
<tr>
<td>Lane change task</td>
<td>Specific lo-fidelity virtual driving environment</td>
<td>Multi-task motorway driving scenario</td>
<td>Primary lateral control of vehicle</td>
<td>Standardized approach, control over variables</td>
<td>Difficult to relate results to interface characteristics</td>
</tr>
<tr>
<td>15 second rule</td>
<td>Laboratory/ statically in car</td>
<td>Secondary task achieved without presence of driving task</td>
<td>Secondary task time (whilst stationary)</td>
<td>Simple approach</td>
<td>Only relates to certain aspects of visual demand</td>
</tr>
<tr>
<td>Keystroke-Level Model (KLM)</td>
<td>Modeling exercise</td>
<td>No user trials take place - models expert performance</td>
<td>Secondary task time (whilst stationary)</td>
<td>Quick/cheap, analysis explains results</td>
<td>Only relates to certain aspects of visual demand</td>
</tr>
<tr>
<td>Extended KLM</td>
<td>Modeling exercise</td>
<td>As for KLM, but with additional assumptions</td>
<td>Visual demand of user-interface</td>
<td>Quick/cheap, analysis explains results</td>
<td>Requires reliability assessments</td>
</tr>
</tbody>
</table>
appropriate to the late stages of the design process where a robust prototype is available. Nevertheless, field trials can be extremely expensive and various ethical and liability considerations must be accounted for. An example of a field trial that was carried out in Sweden concerned drivers’ use of intelligent speed adaptation systems (whereby a vehicle’s speed is automatically kept within the speed limit for the current area. Wallen, Warner & Aberg, 2005).

Road Trials

Drivers take part in a short-term (normally less than one day) focused study using a system in an instrumented car on public roads (occasionally on test tracks). For such trials, a wide range of variables may be measured and analyzed (e.g., visual behavior, workload, vehicle control, subjective preference) depending on the aims of the study. Road trials enable more experimental control than field trials, but are still potentially affected by a wide range of confounding variables (e.g., traffic conditions, weather). Furthermore, such a method remains costly to implement and requires robust protocols to ensure the safety of all concerned. Many road trials are reported in the literature, particularly concerning information and entertainment/productivity oriented systems. For instance, Burnett and Joyner (1997) describe a study which evaluated two different user-interfaces for vehicle navigation systems.

Simulator Trials

Drivers take part in a short-term (normally less than one day) focused study using a system fitted or mocked up within a driving simulator. The faithfulness that a simulator represents the driving task (known as its fidelity) can vary considerably, and configurations range from those with single computer screens and game controller configurations, through to real car cabins with multiple projections and motion systems. An example of a medium fidelity driving simulator is shown in Figure 3.

Driving simulators have become increasingly popular in recent years as a result of reduced hardware and software costs, and potentially offer an extremely cost-effective way of investigating many different design and evaluation issues in a safe and controlled environment (Reed & Green, 1999). Nevertheless, there are two key research issues concerning the use of driving simulators. Firstly, it is well known that individuals can experience symptoms of sickness in driving simulators, manifested as feelings of nausea, dizziness, and headaches. There has been considerable research regarding such sickness in virtual environments, and whilst there is still debate regarding the

Figure 3. Example of a medium fidelity driving simulator
Designing and Evaluating In-Car User-Interfaces

theoretical basis for the phenomenon (see for instance Nichols & Patel, 2002), there is practical guidance for those using driving simulators. For instance, screening questionnaires can be used to eliminate individuals who are most likely to experience sickness during a trial (Kennedy et al., 2001). Furthermore, various countermeasures can be used in the development of the simulator and its environment to reduce the prevalence of sickness (e.g., high, consistent frame rate, air-conditioning, natural background lighting. Nichols & Patel, 2002).

A second and more complex issue concerns validity, particularly behavioral (or construct) validity, that is, the extent to which drivers behave in the simulator as they would in the real world (Blaauw, 1982; Reed & Green, 1999). Driving simulator validity is problematic to study for several reasons. Running both road and simulator trials which are comparable (in terms of participants, tasks, measures, procedures, etc.) can be extremely difficult to achieve, and ultimately will be resource intensive. Furthermore, validity in this area is widely recognized to be a function of a large number of variables, including those relating to how the vehicle is represented (e.g., primary and secondary control design, the sense of enclosure, viewing angles, engine noise, vibration, motion, etc.) and those concerning the driving environment (e.g., visual field of view, screen resolution, graphical complexity, traffic representation, wind/road noise, etc.) Kaptein et al., 1996; Peters & Peters, 2002). Most importantly, our understanding of validity must consider the driving task itself. Driving is a complex task, involving a substantial number of discrete physical, perceptual and cognitive behaviors, and a specific simulator configuration will only enable a subset of these to be investigated (e.g., speed control, headway maintenance).

As a consequence, despite the importance of the topic, there are few driving simulator validity studies in the open literature. Moreover, various limitations can be expressed for previous research in this area:

• It is difficult to generalize from existing validity studies, as they tend to be very specific to (a) the simulator configuration under investigation, and (b) the technology (hardware and software) available at that time (see for instance, Tornros (1998) compared with Blaauw (1982)).
• Studies inevitably only concern a small number of variables, for instance the effect of screen resolution and forward field of view on speed and headway choice (Jamson, 2001); or the effect of characteristics of torque feedback for steering on curve negotiation (Toffin et al., 2003).
• Studies often do not report critical data regarding the simulator configuration (e.g., steering sensitivity, max/min acceleration, driver eye height) which, for given types of study will be critical in results interpretation and cross-study comparison.

Occlusion

This is a laboratory-based method which focuses on the visual demand of in-vehicle systems. Participants carry out tasks with an in-vehicle system (stationary within a vehicle or vehicle mock up) whilst wearing computer-controlled goggles with LCDs as lenses which can open and shut in a precise manner (see Figure 4). Consequently, by stipulating a cycle of vision for a short period of time (e.g., 1.5 seconds), followed by an occlusion interval (e.g., 1.5 seconds), glancing behaviour is mimicked in a controlled fashion. Occlusion offers a relatively simple method of predicting visual demand, but is has been pointed out that its emphasis on user trials and performance data means that it requires a robust prototype and is therefore of limited use early in the design process (Pettitt et al., 2006).
Following considerable research, the occlusion method has recently been formalized as an international standard (ISO, 2005). In particular, guidance is given on how many participants are required, how much training to give, how many task variations to set, data analysis procedures, and so on. Moreover, two key metrics are stipulated: total shutter open time (the total time required to carry out tasks when vision is available); and resumability (the ratio of total shutter open time to task time when full vision is provided). For resumability, there is considerable debate regarding the merit of the measure. Advocates believe the metric provides an indication of the ease by which a task can be resumed following a period without vision (Baumann et al., 2004). Critics point out that the metric is also influenced by the degree to which participants are able to achieve tasks during occluded (non-vision) periods (Petitt et al., 2006). Consequently, it can be difficult for a design team to interpret the results of an occlusion trial.

**Peripheral Detection Task**

This method requires drivers to carry out tasks with an in-car system (either on road or in a simulator) and to respond to the presence of lights within their periphery. The speed and accuracy of responses are considered to relate to the mental workload and distraction associated with secondary tasks (Young et al., 2003). The advantage of this method over occlusion is that it offers an assessment of cognitive, as well as visual demand (of relevance to the assessment of speech interfaces, for instance). The primary disadvantage is that the method still requires some form of driving task. Moreover, in contrast with occlusion, the method has not been fully standardized, and the ability to make cross study comparisons is severely limited by the specific choice of driving task scenarios (affecting task load and the conspicuity of the peripheral stimuli). It has also been noted that it is very difficult to discern between the level of cognitive demand and the visual demand for a given user-interface (Young et al., 2003).

An interesting recent development addresses some of these limitations. Engstrom, Aberg and Johansson (2005) considered the potential for the use of a haptic peripheral detection task, where drivers respond to vibro-tactile stimulation through the wrist whilst interacting with an in-vehicle system. Clearly, such a variation of peripheral detection is not affected by variations in lighting conditions. Furthermore, the authors argue on the basis of their validation work that this method provides “a ‘pure’ measure of cognitive load not mixed up with the effect of simply looking away” (p.233).
Lane Change Task

This method occurs in a basic PC simulated environment in which drivers are requested to make various lane change maneuvers whilst engaging with an in-vehicle system. The extent to which the profile of maneuver made by a driver varies from the optimum maneuver (the normative model) is considered to be a measure of the quality of their driving. Specifically, the method has the ability to assess the impact of an in-car computing system on a driver’s awareness of the driving environment (perception, reaction), and, their ability to safely control the vehicle (maneuvering, lane keeping) Mattes (2003). Considerable research is ongoing with the lane change task in an attempt to develop an international standard (Transport Canada, 2006). Key research issues concern participant choice, training requirements and developing acceptable limits for performance.

15 Second Rule

Participants carry out tasks with an in-car computing system whilst stationary within a vehicle or mock up (i.e., with no driving task) and with full vision. The mean time to undertake a task is considered to be a basic measure of how demanding visually it is likely to be when driving (Green, 1999). A “cut-off” of 15 seconds has been set by the Society for Automotive Engineers (SAE). If the task on average takes longer than 15 seconds to achieve when stationary, it should not be allowed in a moving vehicle. The method is simple to implement and has the key advantage that it has been formalized in an SAE statement of best practice (SAE, 2000).

Research by Green (1999) and other research teams (e.g., Pettitt et al., 2006) has shown strong correlations between static task times and the total amount of time spent looking away from the road at displays/controls, both in simulator and road studies. However, the correlation between static task times and the duration of single glances towards an in-vehicle display is generally poor. This is important because a user-interface may promote a small number of very long glances (e.g., as a result of dynamically changing visual information) which can have a considerable negative effect on driving performance (Burnett & Joyner, 1997). It is for this primary reason that many authors advocate the use of the occlusion method as a better low-cost method for investigating the visual demand of an in-car user-interface (Pettitt et al., 2006; Stevens et al., 2004).

Keystroke Level Model (KLM)

The KLM method from the GOMs family of techniques is well known to HCI researchers and (to a lesser extent) practitioners (Preece et al., 2002; Shneiderman, 1998). It is a form of task analysis in which system tasks with a given user-interface are broken down into their underlying physical and mental operators, e.g., pressing buttons, moving hand between controls, scanning for information. This is a method that is extremely cheap to implement, as there is no need for participants, and the method can be used with very basic prototypes early in the design process. Time values are associated with each operator and summed to give a prediction of task times. Researchers have developed new operator values relevant to the in-car situation (e.g., time to search a visual display, locate a control, move hand back to steering wheel) and have reported strong correlations between predicted task times and times based on user trials (Green, 2003; Pettitt et al., 2005). Task times can be related to certain measures of visual demand for in-car user-interfaces.

In an extension of the KLM method, Pettitt, Burnett and Stevens (2007) recently developed new rules that enable designers to develop predictions for a broader range of visual demand measures. In particular, the extended KLM considers a time-line view of an interaction in which a cycle of vision/non-vision occurs with a user-interface (similar to the occlusion protocol).
The authors have found that their version of KLM can differentiate between tasks as effectively as does the occlusion technique, but recommend that further development is carried out to ensure that practitioners can utilize the method reliably.

CASE STUDY: VEHICLE NAVIGATION SYSTEMS

To ground many of the issues previously mentioned, a specific system type has been chosen for further discussion (vehicle navigation systems). Many of the individual points made for this system can be generalized and are applicable to other in-car computing technologies.

Vehicle navigation systems aim to support the strategic (e.g., route planning) and tactical (e.g., route following) components of the overall driving task. They have the greatest potential to assist drivers who undertake many unfamiliar journeys, for instance as part of work, or during leisure trips (e.g., when on holiday) and those who experience extreme difficulties with existing methods of navigation (particularly paper maps). When linked with reliable, real-time traffic information (thus providing dynamic guidance), the perceived utility of navigation systems to the everyday motorist is significantly enhanced (Bishop, 2005).

The market potential for vehicle navigation systems has already been demonstrated in Japan, where the technology has been available since the early 1990s. Approximately 40 percent of all vehicles on Japan’s roads now have a navigation system installed (http://www.jetro.go.jp/en/market/trend/topic/2004_12_carnavi.html, September 2006). In many other countries, the popularity of navigation systems is currently reduced in relation to Japan, but is predicted to rise rapidly over the next few years (Bishop, 2005).

The majority of human factors issues relevant to this form of technology relate to overload, although as shall be seen, underload is increasingly being researched. With respect to overload, clearly, a key concern is the potential for driver distraction and there has been considerable research on this topic since the mid 1980s (see Young et al., 2003 and Srinivisan, 1999, for reviews). In using a vehicle navigation system, drivers must interact with controls (e.g., to enter a destination, change map scale) and view/understand displays (e.g., to decide which turn to make, to examine options within a menu). In many cases, these interactions will arise when the vehicle is in motion. Consequently, to provide guidance for designers, researchers have aimed to understand how the user-interface design for a vehicle navigation system impacts on both navigating and primary driving performance. Specifically, research has aimed to answer the following three design-oriented questions:

What Information Should a Navigation System Provide?

To support route following, there are a wide range of different information types that a system could present, either referring to something real in the road environment (junction representations, street/road signs, landmarks, etc.) or indirectly referring to or pointing at aspects of the environment (distance to turn, directions, etc.). In this respect, researchers have established through a range of methodologies that the use of distinctive features of the environment (landmarks) within navigation instructions (e.g., “turn right at the church”) offer considerable advantages over the use of distance to turn information (e.g., “turn right in 300 meters” Burnett, 2000; Ross, May & Grimsley, 2004). Moreover, research has identified the fundamental characteristics of landmarks which designers of vehicle navigation systems and providers of underlying map databases must consider in choosing appropriate landmarks for presentation by a navigation system (Burnett, Smith & May, 2001).
How Should Information be Presented?

Navigation and related information has to be presented to the driver in some way, and there has been considerable research on a range of topics. One key concern has been the impact of system modality (voice and/or visual) on driving and navigating performance. The general consensus here is the primary modality for presentation of navigation instructions should be auditory to reduce the conflict with the predominately visual driving task. However, information should also be presented visually, in particular, to support driver’s understanding of more spatially complex maneuvers which cannot be represented easily in voice directions (Ross et al., 1995). Recently, Van Erp (2005) investigated empirically the potential for the use of passive touch as a novel modality for presentation of navigation instructions (specifically, vibro-tactile direction and distance to turn presented through the driver’s seat). They concluded that haptic navigation displays offer various advantages over visual displays, for example, they provide a ‘private’ display to the driver appropriate for very simple maneuvers. Nevertheless, it must be noted that the authors did not make comparisons with the prevailing visual and auditory interfaces. Other research related to information presentation has considered a wide range of issues, such as the format of information (map-based vs. turn-by-turn based), the scheduling of information (when to present instructions), and the location of information (positioning of displays). On these topics, the reader is directed to Ross et al. (1995) and Srinivisan (1999).

How Should Drivers Interact with a Navigation System?

For drivers (or passengers) to interact with a vehicle navigation system, there must be a means by which they can enter data (e.g., postcode for an address), select from continuous/discrete options (e.g., voice volume levels, stored destinations), request/repeat information (e.g., voice directions), and move through the system (e.g., within and between menu screens). There is understandably a natural tendency for designers to utilise the familiar desktop computing paradigms, thus utilizing specific hardware devices (e.g., joysticks, touchscreens, buttons) and associated software approaches (e.g., use of menus, lists, scrolling). Historically, such paradigms were conceived as a means of overcoming the significant limitations of command-line user-interfaces and provided a what-you-see-is-what-you-get (WYSIWYG) experience for the user (Shneiderman, 1998). In the driving context, several studies have shown that such highly visual-manual user-interfaces can have a considerable impact on safety (Nowakowski, Utsui & Green, 2000; Tijerina, Palmer & Goodman, 1998).

As an alternative to such user-interfaces, speech shows promise as a largely non-visual/manual input method for navigation systems (Tsimhoni, Smith, & Green, 2002). Nevertheless, research has also shown that there is considerable potential for cognitive distraction with speech interfaces (Gärtner, König, & Wittig, 2001), and it is critical that recognition accuracy is very high. Moreover, designers must provide clear dialogue structures, familiar vocabulary, strong feedback and error recovery strategies. These issues are of particular importance given the potentially large number of terms (e.g., towns, street names) that might be uttered and the difficulties that a speech recognition system can experience with alphabet spelling (specifically, the ‘e-set’—b, c, d, e, g etc.).

Recent research has also shown the potential for handwriting recognition in a driving context for inputting alphanumerical data (Burnett et al., 2005; Kamp et al., 2001). Whilst handwriting requires manual input, there is a reduced cognitive component and it is a more familiar method for users in contrast with speech interfaces. Nevertheless, issues relating to recognition accuracy remain
and it is critical to place a handwriting touchpad in a location that facilitates the use of a driver’s preferred hand (Burnett et al., 2005).

The difficulties for complex interactions with vehicle navigation systems are considered to be so significant that many authors believe that systems should disable “overly demanding” functionality when the vehicle is in motion (e.g., by “greying out” options when the vehicle is moving. Burnett, Summerskill & Porter, 2004; Green, 2003). This is currently a rich area for research, requiring an understanding of (a) what is meant by “overly demanding,” (b) establishing valid/reliable metrics for the assessment of demand and finally, (c) deciding where to put limits on acceptability (Burnett et al., 2004).

**Underload for Vehicle Navigation Systems**

In contrast with the overload perspective, over the last five years some researchers have viewed navigation systems as a form of automation, where underload issues become central. Vehicle navigation systems calculate a route for a driver according to pre-defined algorithms and then present filtered information, often via visual and auditory instructions. Two related concerns are emerging as important research questions, of particular relevance to user-interfaces which place a reliance on turn-by-turn guidance.

Firstly, it has been noted that there may be a poor calibration in the perceived versus objective reliability of in-car computing systems (Lee & See, 2004). This is of relevance as a navigation system (particularly the underlying digital map) that is unlikely ever to be 100 percent reliable. Nevertheless, drivers, largely based on their accumulated experience, may believe this to be the case. In certain situations, such overtrust in a system (commonly referred to as complacency) may lead to drivers following inappropriate routes and potentially making dangerous decisions, for instance, turning the wrong way down a one-way street. There is plenty of anecdotal evidence for such behavior in the popular press (e.g., http://www.timesonline.co.uk/article/0,,2-2142179,00.html, September 2006). Recently, research has replicated the effect in a simulated environment and indicated that there are considerable individual differences in the likelihood of a driver showing a complacency effect (Forbes & Burnett, 2007). Further research is considering what role the extended user-interface (training procedures, manuals, marketing information) can have in reducing complacency effects.

Secondly, drivers who use vehicle navigation systems may not develop a strong mental representation of the environments in which they travel, commonly referred to as a cognitive map. It has been stressed that traditional methods (e.g., using a paper map) require drivers to be active in the navigation task (route planning and following. Jackson, 1998; Burnett and Lee, 2005). Whilst the demands (particularly the cognitive demands) can initially be high, drivers who are engaged are able to develop landmark, then route knowledge, ultimately progressing to a map-like mental understanding (survey knowledge). Such a well-developed cognitive map means that drivers are able to navigate independent of any external source of information. Empirical research in this area has shown that drivers using current forms of user-interface for vehicle navigation system do indeed experience reduced environmental knowledge in relation to drivers using traditional methods (Burnett & Lee, 2005; Jackson, 1998). A key research question here is how user-interfaces can be developed which balance the need for low demands (workload) whilst simultaneously aiding drivers in developing a well formed cognitive map (Burnett & Lee, 2005).

**FUTURE TRENDS AND CONCLUSION**

The incessant growth in the use of cars and worries about road safety have led car manufacturers...
Designing and Evaluating In-Car User-Interfaces

to offer more intelligent cars providing a range of novel functions to drivers. Moreover, existing mobile technologies such as PDAs, MP3 players, mobile phones, and so on, are increasingly being used within cars, as drivers seek to be more productive and to enjoy the time spent in their vehicles.

All of these computing-based systems offer potential benefits to drivers. This chapter has focused on some key design issues for user-interfaces from the perspective of the individual driver. However, as systems become commonplace within vehicles, there are fundamental conflicts to resolve between the requirements of an individual versus the overall traffic system. In this respect, the design of an in-car computing user-interface will be a critical consideration. As an example scenario, one can envisage many drivers using information systems providing the same information at the same time. Such a situation may lead to a range of problems, for instance the use of roads not designed for high volumes of traffic. Clearly, there is a need for overall management and an understanding of the impact that specific styles of user-interface will have on driver behavior.

A second broad issue for research concerns the interaction between multiple systems. This chapter has introduced the overload and underload concepts and discussed them in turn relating them to different individual systems. It is highly likely that in the short to medium term, overload will be given a prominent position in research and development work, whereas underload will emerge as an increasingly important topic in the medium to long term. However, this singular view neglects the fact that information and control-based systems are likely to be used together in a vehicle. Clearly, there will be various interaction effects for researchers to investigate. Moreover, there is a fundamental need to find the right balance between the two extremes of overload and underload. As noted by Dewar (2002, p. 330), “humans operate best at an optimal level of arousal, and either too much or too little workload can be detrimental to performance.”

The development of suitable methods for designing and evaluating in-car computing user-interfaces will continue to be an important research topic. Reliable and valid methods are required which are accepted within industry. A key motivation will be to establish ‘quick and dirty’ methods (and associated metrics) enabling designers to understand the likely demands of their user-interfaces early in the design process when very rudimentary prototypes are available. A further critical requirement is for “benchmarking,” that is, establishing a point of reference from which user-interfaces can be compared or assessed. Such benchmarks will be of particular benefit when identifying user-interface designs that are considered acceptable or unacceptable, particularly from a safety perspective.

REFERENCES


Designing and Evaluating In-Car User-Interfaces

conference (Human-Centred Computing), Vol. 3, (pp. 180-184).


ISO. (2005). Road vehicles—Ergonomic aspects of transport information and control systems—Oclusion method to assess visual distraction due
Designing and Evaluating In-Car User-Interfaces


Designing and Evaluating In-Car User-Interfaces

Taylor & Francis, Contemporary Ergonomics 2006 (pp. 219-223).


**KEY TERMS**

**Driver Distraction:** Occurs when there is a delay by the driver in the recognition of information necessary to safely maintain the lateral and longitudinal control of the vehicle. Distraction may arise due to some event, activity, object or person, within or outside the vehicle that compels or tends to induce the driver’s shifting attention away from fundamental driving tasks. Distraction may compromise the driver’s auditory, biomechanical, cognitive or visual faculties, or combinations thereof (Pettitt & Burnett, 2005).

**Driving Simulators:** Provide a safe, controlled and cost-effective virtual environment in which research and training issues related to driving can be considered. Simulators vary considerably in their fidelity (i.e., the extent to which they replicate aspects of real driving).

**In-Car Computing Systems:** Provide information to support the driving task or control some aspect/s of the driving task. In-car computing systems may also provide information and/or services that are unrelated to driving.

**Keystroke Level Model:** Is an established HCI method used to predict expert’s task times with a user-interface. It can be used with in-car user-interfaces to predict static task time, that is, the time taken to achieve tasks in a stationary vehicle. Recently, the KLM has been extended to predict visual demand measures related to the occlusion protocol.

**Overload:** (Due to in-car computing systems) occurs when a driver’s information processing resources are overwhelmed and performance on primary driving tasks inevitably suffers.

**Underload:** (Due to in-car computing systems) occurs when automation of core driving tasks (such as steering, braking, etc.) has led to a situation in which driving performance has deteriorated. This may have arisen because the driver has reduced awareness of other road users, has changed their behavior in negative ways or has inferior skills/knowledge in driving.

**The Occlusion Protocol:** Is a user trial method used in the design and evaluation of in-car user-interfaces. Participants typically wear LCD glasses which restrict the visual experience by only enabling short (e.g., 1.5 seconds) chunks of visual attention with an in-car user interface. Measures related to the visual demand of an interface can be established.

This work was previously published in *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, edited by J. Lumsden, pp. 218-236, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 2.12
The Semiotics of Smart Appliances and Pervasive Computing

Peter Bøgh Andersen
University of Aarhus, Denmark

Martin Brynskov
Center for Interactive Spaces, University of Aarhus, Denmark

ABSTRACT

This chapter presents digital habitats, a conceptual and methodological framework for analyzing and designing smart appliances in the context of pervasive computing. The concrete topic is a project in pervasive gaming for children. The framework consists of a set of theoretical concepts supplemented by diagrams for representing semiformal models. We give a short overview of selected theories of play and gaming and apply the framework to an implemented simple pervasive game. Finally, we use the framework in a constructive manner to produce a concrete design of a new game. The result is discussed and compared to other approaches. The main points are the following: (a) it can describe communicative as well as material acts plus the way they hang together; (b) it provides an explicit link between human activities and their spatial context; (c) it has an explicit dynamic model that precisely describes the conditions for executing actions; and (d) it offers a typology of participant roles based on linguistic theory, which supports design processes.

INTRODUCTION

In this chapter, we will present an approach to analysis and design of computing systems that transcends the boundaries of traditional office computer systems such as PCs and laptops. These transcending systems are called ambient, ubiquitous, or pervasive computing systems, and they pose new challenges to the way we understand, analyze, and design information technology. With such systems, computing power spreads from dedicated computing hardware into other artifacts and places, both at the workplace and
in everyday life. Microcontrollers, sensors, and actuators have been embedded in machines for decades, but the functionality was tied closely to the artefact in which it was embedded (e.g., a washing machine or the side mirror of a car), and therefore, the computational “smartness” was not foregrounded in itself. Two factors have changed this: (a) the increasing flexibility and computing power of smaller-scale devices, and (b) the wireless networking capabilities and structured exchange of information. In this world of smartphones, GPS (location tracking), and software agents, we need concepts to communicate about human needs and activities as well as technical infrastructures. Digital habitats is a suggestion for such a framework. In our presentation, we focus on fun and games, but the framework originally was conceived in a workplace setting.

“On the Concept of Intelligence” discusses the concept of intelligence and concludes that the everyday use of the concept applies to a disposition to act in a certain way in a network of other actors and artefacts. To be intelligent is to fill out Agent roles appropriately in activities conducted by such networks.

The habitat concept defines the framework we use. On the one hand, it defines activities in terms of roles, participants, actions, and glue binding participants to roles. The well-known automatic door opener is used as an example. On the other hand, it offers a maplike representation that ties activities to physical and informational spaces and describes the various types of interplay between physical and informational space. This is particularly relevant to pervasive computing, since pervasive computing is characterized by being distributed in physical space and by overlaying physical space with digital information.

In “Play and Games,” we give a short overview of theories of play and gaming. In “Designing Pervasive Games,” we adapt the framework to the domain of games and describe an implemented simple pervasive game called StarCatcher. “The Bogeyman” puts all the ends together in a concrete design of the game Bogeyman that elaborates on StarCatcher by drawing on the presented theory. “Related Work” compares this chapter to related fields, and the conclusion summarizes the advantages of the present approach, as we see it.

**ON THE CONCEPT OF INTELLIGENCE**

The first issue that must be discussed is what *smart* and *intelligent* mean.

**Intelligence**

Since there is no universally accepted definition of intelligence (Roth & Dicke, 2005), we accept Gilbert Ryle’s (1970) claim that these words denote (a) the manner in which an action is performed and (b) a prediction about the way other actions are performed. In this case, intelligent does not denote a special mental process that is the cause of the action but rather a disposition generally to act in a certain manner. What is intelligent depends upon the circumstances but often involves features such as: the action achieves its goal, it does not contain superfluous steps, it economizes with resources, it does not destroy or hurt participants, it is an innovative way of solving a difficult problem, and so forth.

This definition is similar to a prevalent view in contemporary cognitive science, that “mental or behavioral *flexibility* is a good measure of intelligence, resulting in the appearance of novel solutions that are not part of the animal’s normal repertoire” (Roth & Dick, 2005, 250). We choose to focus on the behavioral aspect and, thus, preclude ourselves from making inferences about neuroanatomy or mental mechanisms. On the other hand, this choice to focus strictly on behavior allows us to use the word about humans as well as artifacts without committing ourselves to philosophical doctrines about the nature of the mind (see Dennett, 1991).
The Semiotics of Smart Appliances and Pervasive Computing

If intelligence denotes a disposition to act in a certain way, then it follows that its unsophisticated and immediate reference is to networks of actors, instruments and objects. The reason is quite simply that human action normally is mediated by an instrument and directed toward an object. These elements form a network, and it is such networks that we can immediately call intelligent or smart. There will often be a next analytical step in which we ascribe the main honor to one or more of the participants in the network: one of the actors, tools, or objects may be seen as the main contributor.

Networks of Stupidity

This analysis is easier to verify with negative predicates such as negligence, stupidity, and inability. For example, accident reports must point to a participant that is guilty of the accident in order to suggest future remedies against the type of accident and because of insurance issues. Although accidents are mostly caused by a particular configuration of participants in the network, the report must point out the weak link in the chain. However, this is often difficult and a matter of interpretation.

Here is an example, due to PhD student Thomas Koester: a ferry was fitted with a system that compensated the heeling of the boat by moving water in the ballast tanks. The system had a manual and an automatic mode, but sometimes it would unexpectedly go from automatic to manual mode. The mode change was displayed on the bridge and in a closed locker on the deck. The accident occurred when the deck crew was emptying the deck for its cargo of trucks and cars. When cargo is removed from one side of the ship, it will heel, and the system is supposed to compensate; in this case, it had switched to manual. The result was that the heeling was more than six degrees, and the ramp was damaged. Who was to blame? The crew? It knew about the fault; should they have opened the locker and checked the mode regularly? But then the disembarkation would have gone more slowly. The manufacturer? The system obviously should not be allowed to switch to manual by itself. The ship owner? He could have counteracted the fault by mounting a warning system on the deck, informing the deck crew of the mode shift.

The fault is clearly due to the whole network: a faulty system + missing indications of the error + lack of attention in the crew. It is a practical and political matter, not a philosophical issue, to decide on a remedy: correcting the mode error, mounting a warning light on the deck, or changing the operation procedures.

The importance of networks vis à vis individual actors has been emphasized in the actor network theory (Latour, 1994, 1999; Law, 1987). For example, power is a predicate that pertains to an actor’s position in a network, not to the individual actor.

The example illustrates the point made previously: predicates such as intelligent, smart, sloppy, inefficient, and so forth are in the first place predicates of networks, not of their parts. Only analytically can one of its participants be singled out as the main factor, and this choice depends upon the countermeasures that are considered feasible and desirable.

The same line of reasoning can be applied to intelligence. The IBM chess program Deep Blue beat grandmaster Kasparov on May 4, 1997. Therefore, there is an intelligent network of actors comprising at least the following participants: the developers, the chess literature they used, the system, and the operator. But which participant should be picked as the winner of the match? It is as difficult to decide as in the accident case, but it was the development team behind Deep Blue that took home the $700,000 first prize. Thus, in practice, the development team was singled out as the intelligent part of the network.
Intelligent Technology

If we are to single out one participant of an intelligently conducted activity, it must be because its contribution is particularly conspicuous. How do we decide this? One way is a simple substitution: if we keep the chain constant and replace one participant, does the performance of the chain become more or less intelligent? If the chain performs less intelligently, we may tentatively attribute intelligence to this part of the chain.

From these arguments follows a definition of smart/intelligent technology:

1. Intelligent technology is a kind of technology that is able to contribute positively to activities whose manner of performance we will intuitively call intelligent.

   In the following, we incorporate this definition in a broader framework, which we call a habitat. The habitat concept is a general attempt to link networks of activities and actors to physical space on the one hand and signs and sign usage on the other hand.

THE HABITAT CONCEPT

The habitat concept was elaborated in the research group Frameworks for Understanding Software Systems (Andersen & Nowack, 2004; Brynskov & Andersen, 2004; May & Kristensen, 2003). The purpose of the research was to devise a methodology that is better suited for handling context-sensitive pervasive computing than are traditional ones. Traditional object-oriented modeling methods (Mathiassen, Munk-Madsen, Nielsen, & Stage, 2001) are characterized by two features: (a) they model conceptual structures while the physical system represented by these models is absent in the method, and (b) the physical location of the software components is postponed to the last implementation phase (Andersen & Nowack, 2002). This is a problem in context-sensitive pervasive computing, because one of its possibilities is the use sensors to collect information about physical objects located in the spatial or temporal vicinity of the physical device where the software resides. Therefore, we need to coin concepts for the information available in the vicinity of the physical device and for the physical objects referred to by its software model. The former concept is called the access area; it denotes the locations from where sensors can access information; the latter is called the reference area and denotes the location of the objects referred to by the software model. Furthermore, a description of information is incomplete if we do not specify its use (i.e., what activities is the information used for?). Therefore, we need to enter the notion of activities into our framework. Finally, we are interested in systems that are sensitive to their physical surroundings, which means that our third component must be space. We are interested in physical spaces that are designed or have evolved to support a delimited set of human activities. This is true for an overwhelming number of the spaces in which we live daily: private houses, airports, hospitals, railroad stations, road networks, and so forth. We have chosen the term habitat to denote a physical space that is designed to support some set of activities and that provides access to information about objects relevant to the activities (Brynskov & Andersen, 2004).

2. A habitat is a chunk of space-time that is designed or has evolved to support a delimited set of activities by offering physical artifacts and information sources useful for conducting the activities.

   From this definition, it follows that a habitat should be characterized along three dimensions:

3. The physical habitat: The physical layout and boundaries of the habitat plus the available physical artifacts.
4. **The informational habitat:** The signs available (access and reference area) to participants in the activities (digital and nondigital signs).

5. **The pragmatic habitat:** The action affordances offered by the habitat, the macro-roles, and the role-requirements of the participants.

The contribution of the present chapter is threefold: (a) it explores the usefulness of the habitat concept in the domain of pervasive games; (b) it presents a framework for describing communicative and material activities and the way they interact; and (c) it presents a notion of intelligent or smart technology that is consistent with the framework. The chapter mostly draws on semiotic and linguistic theory, and our basic understanding of the habitat concept is well-captured in Peirce’s semiotic triangle (Figure 1). The space itself and its manufactured representations (e.g., signposts and electronic displays) are representamens; the interpretant of these signs is the activities associated to the habitat, and the object is the phenomena inside the reference area (i.e., things or events that are relevant to the activities).

**Activities**

In order to use definition (1) for design purposes, we need a general functional definition of activities, and a set of context-dependent definitions of what criteria intelligent activities must meet in various domains. The general definition is given in this section. The game-specific adaptation is presented in the section “Applying the Theory to Games.”

**Goal, Roles, Participants, and Actions**

As in the case of intelligence, we adapt a functional view of activities (Andersen, 2004a, 2004b, 2005; Bødker & Andersen, 2005; Brynskov & Andersen, 2004). The conceptual framework consists of roles, participants, actions, and activities. Activities consist of actions subsumed under a shared goal, and participants play roles in relation to actions and activities.

At the action level, we use traditional linguistic semantic roles like agent, theme, instrument, beneficiary, source, destination, time, place, direction, and so forth (see Fillmore, 1968, 1977; Blake, 2001). At the activity level, we use macro roles like librarian/borrower, customer/clerk, lawyer/client, judge/defendant, and so forth. The macro roles are defined by the semantic roles that participants can play in actions: thus, a Judge is the Agent of actions like planning the sessions of the court, considering the judgment, giving a verdict, sentencing, and so forth. This is the method used in the classical work on narratology (Propp, 1968).

In addition to specifying actions, roles, and activities, we need to specify the goal of the activity. We distinguish between four types: creating, destroying, preventing, and maintaining a state (Lind, 1994).

Not all participants are equally qualified to play these roles; a role requires its filler to possess the requisite incentives and qualifications in order for it to contribute in an intelligent and competent manner. Incentives include intentions, desires, and obligations; qualifications cover the relevant abilities, rights, and knowledge. Sometimes we
will refer to these qualifications as a lump of glue that binds a participant to a role.

Some of these requirements are realized differently in humans and nonhumans. For example, intentions are implemented as negative feedback loops in automatic systems like thermostats that aim at keeping a certain process variable (e.g., the heat) locked at a certain set-point. Its goal is to maintain the process variable at the set-point. In software agents (Russell & Norvig, 2003), intentions often are implemented as some kind of means/end structure along with a planning and execution algorithm. The goal often will be to create a certain state, but in order to do so, the algorithm will have to maintain other states (protected states) and prevent others (forbidden states).

According to the approach outlined previously, we should not necessarily ask whether this implementation resembles the way humans make and execute plans. Instead, we should ask for qualifications: to which degree is the behavior of the automatic system suited for playing its assigned roles in the activity? The same goes for knowledge: there are many types of knowledge representations (predicate calculus, semantic networks, frame systems, conceptual graphs, and neural networks), but for the present purpose, it is not interesting whether humans use something comparable but only to which degree the system component using these methods is able to play its allotted role. The viewpoint, thus, is strictly behavioral.

In this chapter, we will concentrate on cooperative activities involving humans, intelligent systems, and dumb artefacts. Since the activity is cooperative, the following requirements of intelligent agents are important:

6. The agent should have intentions to act that are understandable to the others (comprehensibility);
7. Agents should display their intentions and actions in a way that is accessible to the other participants (accessibility);
8. Agents should be able to perceive and manipulate the attention-focus of their colleagues and adapt their own actions accordingly (joint attention);
9. Agents should be able to enter and leave the activity (intermittent participation).

Applied to mechanical agents, comprehensibility (6) means that they only should form plans that are understandable to their human and nonhuman colleagues. Understandable algorithms have a higher priority than cunning ones, but there is a tradeoff between intelligibility and efficiency.

The intentions of the agent should be signaled to the other participants in a way that allows them to reliably infer intentions from behavior (accessibility, 7). In other words, there is a mutual commitment to signal the truth, the whole truth, and nothing but the truth in the least complicated manner (equivalent to the so-called Gricean maxims [Grice, 1975] from the pragmatics of natural language). On the other hand, once the deliberate representation of intentions has been introduced, it is possible to misuse it to lie about intentions as well.

Attention is intentionally directed perception (Gibson & Rader, 1979; Tomasello, 1995). Joint or shared attention (8) is necessary for cooperation: If I cannot see what my colleague is concentrating on, I cannot collaborate with him or her. However, shared attention is more than just looking at the same object or detecting gaze direction; it is the mutual understanding between two intentional agents that they are, indeed, intentional agents (Tomasello, 1995) (see Dennett’s, 1987, The Intentional Stance). In order for me to share a goal-directed activity with somebody, I not only must be sure that the other participants have their minds focused on the objects or topics I think of myself, but all participants also must share the fundamental assumption that activities can be goal-directed and that they can be coordinated by detecting and manipulating attention. In fact, shared attention is a constituent factor in cultural
Humans share attention all the time, beginning near the end of their first year of life as a result of a constellation of emerging social behaviors (e.g., communicative gestures, imitation, social referencing) (see Tomasello, 1995). We can read or infer and, therefore, continuously follow and manipulate each other’s focus of attention (and thereby intention) throughout a joint activity. This interpreted focus is a common anchor point without which two persons would drift away from each other—mentally and physically—instead of staying focused on the same subject matter. They would be solipsist nomads.

Intermittent participation (9) means that participants must be able to leave and enter the network without its breaking down. Much pervasive technology, including games, must allow participants to enter and leave without notice. For example, a door opener would be unusable if the pedestrian were to type in complicated commands preparing the door for the event before walking through the door and signing off after the passage. The ferry system in Networks of Stupidity was not sufficiently robust to allow the mechanical agent to leave the network without notice; the network broke down and damaged the ramp.

Joint attention and intermittent participation are also important in our concrete topic: children’s games. When children play, it is important for them to keep track of the other’s focus. Children gradually develop this sense of playing or doing things together as opposed to just doing things in parallel. At the age of two, they have all the prerequisites for joint attention (Kaplan & Hafner, 2004), and they continue to develop still more complex social interaction based on this fundamental competence. Also, children enter and leave activities very often (i.e., their participation is intermittent and highly dependent on the environment). They have to go and eat, must go outside, or one participant is picked up by his parents.

An example of a game that supports intermittent participation is LEGO Star Wars (www.lego.com/starwars), which allows players to join and leave a game in progress so that a parent can step in and help a child but leave again without having to play the entire game. When a player leaves, the character continues as an independent agent controlled by the software (at least one player must be controlled by a human, however; otherwise, the game ends). Massively Multiplayer Online Games (MMOGs) (e.g., World of Warcraft [www.worldofwarcraft.com]) are another example of games built around intermittent participation. Mobile MMOGs exist, too, and are called 3MOGs (e.g., Undercover 2: Merc Wars [www.undercover2.com]). Often, children can be seen not only as mobile but also as nomadic, because they are dependent on the resources and partners offered by the environment (see the following section). In this case, the proximate environment has a marked influence on their patterns of play (Brynskov, Christensen, Ludvigsen, Collins, & Grønbæk, 2005).

The problem of sharing attention between humans and machines is that machines do not have the ability to read attention, let alone intentional stance (see, however, Kaplan & Hafner, 2004). In the ferry incident, one of the problems was also the lack of joint attention. The automatic pump system switched to manual operation without monitoring whether this important change was brought to the attention of the crew. The crew, on the other hand, could not monitor the attention of the pump system and, therefore, failed to realize that they were supposed to operate the pumps manually. Thus, intermittent participation seems to require joint attention.

People use language to a large extent to signal their focus of attention. With machines in the loop, this is not possible, not because computers cannot produce or parse speech at all (they can to some extent), but because their underlying representation of attention and states do not map easily onto their human partner’s. This is because
the semantics of a language is not context-free but grounded in experience. Humans seem to take the other’s perspective by mapping observations onto their own nervous system in a kind of simulation. Therefore, joint attention presupposes understandable algorithms.

If we still want humans and machines to share attention during activities, we have two options:

- We can try to develop epigenetic artificial intelligence that can learn to share attention by means of contextual grounding, imitating human cognition (although it is still an open question how exactly to achieve this) (see Steels, 2003; Kaplan & Hafner, 2004), or
- We can try to create a fixed, hardcoded, and more simplistic common ground where both humans and machines can read and interpret behaviors.

We choose the latter, and one of the methods is to design comprehensible algorithms so that there is something for the human participants to share. As a consequence, one might say that the burden of flexibility, to a large extent, is put on the shoulders of the human participants.

The challenge of handling attention focus and intermittent participation is not at all new in human-computer interaction (HCI), especially in the field of computer-supported cooperative work (CSCW) (Bardram & Hansen, 2004). There is a difference, though, in signaling an application’s states and functions as a tool or medium, as is the traditional case in HCI, or a person’s availability in CSCW, to signaling an agent’s states. The difference is whether we attempt to represent intentions (Tomasello, 1995). Attentive agents and calm systems that work in a sort of symbiosis with their owner and sense or guess his or her intentions is an old idea (Licklider, 1960; Weiser & Brown, 1995), and it is still being pursued (Maes, 1994; Maglio & Campbell, 2003). As already mentioned, participants in our framework are glued to their roles not only by intentions but also by desires, abilities, rights, obligations, and knowledge, which also, to a higher or lesser degree, indicate intentions. Therefore, a minimal requirement is to signal these binders not only in a form that humans can interpret but also the other way around. We see play as an interesting case, since this is an important part of the situations in which children learn the intricacies of the semiotics of behavior.

### Door Openers

In this section, we illustrate the concepts used to model activities by a widespread technology; namely door openers. In addition, we offer some simple diagramming techniques. Door openers are good examples because they involve networks of humans and nonhumans, and nonhumans play an Agent role in the sense that they initiate and monitor activities.

The activity concept is the one described previously. We use a diagram consisting of actions, and we highlight the relations between the actions by graphical means. Two actions can be connected by arrows signifying dependencies

![Door Openers](image)

**Figure 2. Automatic door opener in shopping mall**
between participating in the two actions. The rules are as follows:

10. An activity is executed if all participants are able to fulfill their roles to some degree and if the Agent filler is strongly obligated and/or strongly desires to fulfill his or her role. In the game described in “The Bogeyman” section, desires are created by application of dissonance theory (Eskola, 1973), whereas ability is created by the participant’s location. Whenever the necessary participants are assembled in the same place, the activity is enabled.

11. If something successfully participates as role A in executing action X, then its intentions, desire, abilities, rights, obligations, and/or knowledge (its glue) to participate as role B in action Y are changed.

We shall use various ways to diagram these notions (a discussion of diagramming techniques can be found in Andersen, 2004b). Figure 3 shows an ordinary door opener from a shopping mall. The dependency between actions is shown by arrows annotated by the glue dimension affected. The participants involved are shown in boldface.

The activity contains five actions that are connected as follows: when a person walks toward the door in order to get in, the door opener is required to open; when it does, it becomes able to function as a passage for persons passing the doorway but also for heat disappearing. When it closes, it loses these abilities. The goal of the activity is letting people in through the door while preventing heat from disappearing. Since the two goals are contradictory, we must dissolve the contradiction by unfolding it in time. Therefore, when the door opens, it becomes obligated to close after, say, 30 seconds.

In the game-specifications, we use a simplified version of the diagram in Figure 3. The simplified diagram is on a high level of abstraction, displaying only the essence of the game.

The abstraction is made in the following way:

12. The participants, roles, and glue changes responsible for the dependency between the actions are stripped away.

Figure 3. The activity of door-opening focusing on the door participant. Obl = obligation, abil = ability. ‘+’ means increases, ‘-’ means decreases
13. Abilities and rights are merged under the heading *possibility* and represented by a single-headed arrow \(\rightarrow\).

14. Desire and obligation are merged under the heading *necessity* and represented by a double headed arrow \(\leftrightarrow\).

15. Inability or prohibitions are merged as *impossibility* and represented by \(\#\).

The behavior of the door opener now looks as shown in Figure 4.

The reason for these formal abridgements is that diagram design is a tradeoff between precision and overview. The reader should be able to grasp immediately the possible paths of the story by reading the diagram. However, there is still an underlying detailed semantics in which participants, roles, glue, and glue changes are specified. They are just not shown in the diagram.

The previous notations highlight the interdependencies between actions but do not provide a clear picture of the relation between activities and space. Such relations are important when we design context-sensitive, pervasive technology, and therefore, we introduce a maplike notation in the next section. It associates activities to a physical space, if the space plays the role of location in the activity.

**Habitats**

In pervasive computing, the physical space and its relations to the informational space becomes important. Therefore, we need a diagram that highlights the spatial properties of activities and networks, in opposition to the previous section in which we highlighted the relations between actions.

Figure 5 shows a diagram that codes the spatial participants graphically and tones down relations between actions. We have selected the spatial participants (i.e., spaces that participate in the Location role of the activities, spatial participants filling the location roles are called *habitats*. We have shown two habitats: the shopping mall and the entrance. To each habitat is associated the actions that can be performed there: selling and paying in the shopping mall, and walking through doors at the entrance. In addition, we have represented the signs involved in terms of (a) the area from where they can be accessed (the access area) and (b) the area containing the object denoted (the reference area). Both areas are shown by means...
of dashed polygons, and an arrow points from access to reference area. Thus, the door opener has access to the areas inside and outside the door. The arrows are decorated by the main signal path. Sensors often transmit and receive signals in two directions. For example, radar sends out radio waves and records their echo. The outgoing waves are only a means for receiving the ingoing echo. Therefore, the main signal path in radar is from reference to access area, which is also true in the door-opener case.

We distinguish between relative and absolute references. The difference is that if the access area of relative references moves in time or space, the reference area moves, too. This is not the case with absolute references. The reference of the door opener is relative, since if we move the door opener, it will refer to the new environment in which it is placed. Transponders, RFID-tags, and radars create relative references. The distinction also shows up in language, where the so-called deictic words (here, now, me, you) have references relative to the speaker. Compare the two sentences: The Second World War ended in 1945 and The Second World War ends now. The former does not change meaning if it is uttered in year 2000, whereas the latter becomes false if uttered in that year.

To each habitat is associated half-baked actions that can be performed there—the *affordances* (Gibson, 1986) of the habitat. In the following, “#” means that the role is instantiated, whereas a type indication such as pedestrian means that the role can be filled by a specific type of participant. Pedestrian walks through #door thus means that the Location role is instantiated by the particular door and fixed, while the Agent role can be filled by any pedestrian. Fully instantiated actions are created by unifying the action possibilities of the participants with the affordances of the habitat. For example, #smith can walk through doors + pedestrians can walk through #door unifies to #smith can walk through #door, because #smith is a pedestrian and #door is a door. If #smith was not a pedestrian but a car, or if #door was not a door but a wall, unification would be impossible, and no action could be instantiated. There are many ways to implement instantiations. In (10), the “Activities” section, we suggested a particular way of doing it: shared location gives the participants the ability to instantiate roles in the actions, while desire gives a participant the incentive to instantiate the agent role.

This way of combining abilities from agent and environment is useful in agent design, as argued by Cabri, Ferrari, and Zambonelli (2004),

---

**Figure 5. Shopping mall and entrance.** #door means “an instance of the class of"
who also uses a role-based architecture. We shall term it contextual execution. The term means that an execution of an action consists of pieces, some of which originate from the participant, while others come from the habitat. This enables participants to act flexibly in different contexts without having to remember the exact details of each particular context.

The notion of half-baked actions (Andersen, 2004c) reflects the idea that we normally never start from scratch when we act. We know that specific participants normally are useful for the purpose we want to accomplish, and we try them out first. To take the door opener example, if we want to enter a room, we do not start with the isolated verb enter and spend time reflecting which participants it should have. Our point of departure is enter through #door, where the direction case is instantiated by a specific door we can see. Only if the door is locked may we begin considering the windows and the chimney. Another argument is that isolated verbs do not allow predictions: drinking milk, beer, and cyanide has quite different consequences.

In this method, all actions are bound to a certain physical habitat, and, indeed, many actions work in this way. We do need to be inside the airport to embark on an airplane. But what about the planning of the travel? Do we have to be in Portugal in order to plan the travel from the airport to the hotel in Lisbon? No, of course not. Some activities, such as planning, should not be and are not bound to specific locations. Instead, we imagine how things are in Lisbon and simulate the remaining part of the travel in our minds (or on a computer, or let the travel agency take care of the problem). This means that we need the concept of fictive habitats: habitats we imagine and in which we can test solutions without being there in reality. Fiction in general enables us to enjoy living in fictive habitats for a delimited period of time, and therefore, we shall need the concept in “Designing Pervasive Games” and “The Bogeyman,” where we address the question of designing pervasive games for children (on the logic of fictive worlds (see Ryan, 1991). Pervasive games thrive on the interplay between actual and fictive habitats, as we shall see presently.

**PLAY AND GAMES**

The preceding conceptual framework has mostly been used for analytical purposes or for instrumental design topics like brain surgery; evacuation procedures on ships; and pervasive applications in shopping malls, airports, and hospitals (Andersen, 2005; Andersen & Nowack, 2002, 2004; Brynskov & Andersen, 2004; Haase, Musaeus, & Boisen, 2004; Kristensen, 2002, 2003; May & Kristensen, 2003; May, Kristensen, & Nowack, 2001).

Designing technology for playful activities is very different from designing tools for instrumental activities. It is not necessarily the most efficient solution that is the best solution. The user should have fun—and fun is hard to predict and design. Thus, playful activities often are not goal-directed, at least in the sense that the goal may be hard to define or quantify, and the activity may be enjoyable for its own sake (see Csikszentmihalyi’s, 1990, concept of flow).

In the rest of the chapter, we shall test our framework in a design context. In particular, we shall explore whether it yields creative ideas for designing pervasive technology for children’s play. The project called Nomadic Play in Mixed Environments is a part of the Center for Interactive Spaces at the University of Aarhus.

The notion of nomadic play is different from mobile play. Mobile indicates independence of physical location (i.e., activities are enabled, and resources are available regardless of the user’s location). For example, a mobile phone may allow a child riding a bus to play a game, check e-mail, or use instant messaging. Nomadic, on the other hand, indicates dependence of physical location. The analogy is nomads traveling through the desert, being dependent upon oases and family
networks. In a similar fashion, nomadic users are dependent upon, or take advantage of, resources and services that are made available by the environment, including other people’s devices (peer-to-peer). Thus, nomadic play indicates a playful use of smart appliances and pervasive computing systems that could not occur if the child was not at a certain location (e.g., at a club or at home) that offers certain services (e.g., a home game server or media center) to a certain group of people (e.g., friends of the house). One could argue that certain features of mobile systems are also nomadic in nature (e.g., the coverage of mobile phone networks, but we restrict the term nomadic to systems that are more local in the physical sense and, thus, are designed to support a more focused set of human activities.

**Theories of Play and Gaming**

Children’s play has mostly been treated in developmental psychology (Piaget, 1962; Sutton-Smith, 1979; Vygotsky, 1976) and in the context of game design (Salen & Zimmerman, 2004). Here, we adopt the view that play is the search for fun. To be a bit more specific, play is a focused but internally motivated activity without serious consequences. It may not have a goal outside itself other than it keeps a child occupied in a way that allows it to gain experience in the world, both physically and socially. From this perspective, play could be seen (by adults) as training and preparation for adulthood, whether structured or not. The search for fun is the fuel that drives the activity. Thus, learning may be considered play, as long as the internal motivation is high, because it is fun.

Playing a game can be seen as a special case of ludic activities, which, in turn, are a part of being playful (Salen & Zimmerman, 2004). A game can be defined in many ways, and there are plenty of suggestions from which to choose. Koster (2005) lists the following from academia: Roger Callois’ an “activity which is ... voluntary ... uncertain, unproductive, governed by rules, make-believe”;

Johan Huizinga’s “free activity ... outside ‘ordinary’ life ...”; and Jesper Juul’s (2005):

*A game is (1) a rule-based formal system with (2) a variable and quantifiable outcome, where (3) the different outcomes are assigned different values, (4) the player exerts effort in order to influence the outcome, (5) the player feels attached to the outcome, and (6) the consequences of the activity are optional and negotiable.* (Juul, 2005, p. 3)

We distinguish between two types of playful activities: playing and gaming. A game has fixed rules and a quantifiable outcome (a clear goal), whereas play has ad hoc negotiable rules and a fuzzy outcome. A game can be embedded in play (see Juul’s point about the optional and negotiable consequences of a game). Both types of activities should be fun.

An important distinction is that between emergent games and games of progression (Juul, 2005). In the former, the game is defined by a small number of simple rules governing a possibly large number of human or nonhuman participants. An emergent game can be played many times, since each round is different. Emergent games invite planning and strategic thinking. Games of progression are defined by a more or less fixed sequence of events and often tell an elaborate story. Like novels and films, they seldom can be played more than once.

Our two games are games of emergence, but in *Bogeyman*, we have tried to add narrative traits.

**Applying the Theory to Games**

This section adapts the concepts presented previously to the special case of games. In this section, we consider games in general, and in the next section, we focus on pervasive games.

We first show that the glue concept (that binds participants to activities) is useful to characterize games.
The children’s desire motivates them to enter activities (“come and join us; this is fun”) and a lack of desire to leave them (“I am going home because this is boring”). Playthings must create desires in children to assume the Agent role.

*Ability, rights, knowledge and obligations* may regulate participation for human as well as nonhuman participants. An able goal keeper is sure to be enrolled in a soccer game. In the Danish game called The Earth Is Poisonous, it is allowed to tread on furniture but forbidden to step on the floor, because it is poisonous. These tokens of glue may be earned from previous activities, and if they are given a physical shape, they can be exchanged by the children. For example, a child only can participate in secret club activities if he or she has been allowed access to the club and can show his or her membership card.

A child may only get entrance to the club den if he or she knows the secret password. Or abilities may be missing: “You are not allowed to take part because you are stupid.” (One might ask, Who is lacking abilities here?)

Glue like rights and obligations can be distributed: “I’m the president and you must do what I tell you. No, that is silly. Well, then you can be the vice-president and tell Billy what to do.”

Actions can be associated to spaces via the habitat concept. “This is a pirate ship, and the grass is water. You cannot walk on the water. If you do, you drown.” “This chair is a car that can go 200 miles per hour.” “The flagstone is a hole in the Earth, and if you tread on it, you will fall down.”

Games can have different purposes:

- **Creation**: The objective of soccer is to accumulate goals. The goal of exchanging Pokémon cards is to increase one’s collection.
- **Destroying**: The objective of a Star Wars game is to destroy the Death Star. The objective of teasing and mobbing is to destroy the self-confidence of the victim.
- **Preventing**: The objective of a cowboys and Indians play is to prevent the Indians from conquering the fort.
- **Maintaining**: The goal of skipping is to keep the skipping rope moving.

If we focus on pervasive games, the following possibilities present themselves. According to (10), activities can be executed if its role fillers are qualified for the roles and if the Agent filler desires or intends to fill the Agent role. The roles, therefore, are filled by entities that meet the requirements (glue). In the nomadic play scenario, roles are filled by children, artefacts, hardware, software, and locations. Joining and leaving the activity is done on an ad hoc basis, like when children play. At the conceptual level, we do not discriminate between children and software agents. It is the definition of the role at hand that matters. As a consequence, if a software agent meets the requirements to fill a role, it can play it. This also means that a child could fill a role that usually might be played by an artefact or piece of software.

Children and software may compete for filling a role: “I do it” vs. “My mobile phone does it.” More interesting, however, is in-between situations in which child and software together fill a role. Part of the role is automated; part of it is manual. Sometimes a child may be represented by a less capable but sufficient software agent, and agency could be transferred back and forth. Such shifts between manual and automatic control are very common in process control; sometimes the officer of the watch maintains the course, and sometimes the autopilot does it.

A collection of roles that typically go together can be collected in macro roles. These macro roles either could be statically defined, like in the rules of a game, or they could be assigned dynamically, based on the actual history of actions taken by or experienced by the entity (bottom-up role filling). This goes for human as well as nonhuman participants. The latter gives
us interesting opportunities, since the playthings can have a memory that remembers the glue they have earned in previous plays. This is well-known in role-playing games (RPGs). You must earn experience points to become a level-32 wizard. In the extended case, the character can develop behavioral traits and reputation in an emergent fashion, based on previous actions (e.g., by choosing to practice certain skills).

Sharing things is a central part of many activities. We want to be able to share passive objects as well as active or intelligent ones. This also means that pieces of runnable code should be able to migrate live on a variety of different hardware (contextual execution).

The effect is that we get a network, as described previously in actor networks, in which the possibilities are not tied to the entities themselves but rather to the possible roles they can fill in a given set of activities.

DESIGNING PERVERSIVE GAMES

In this and the next section, we use the contextualized concepts previously defined for designing two pervasive games.

The Technology

In the Nomadic Play project, we try to invent new ways of having fun or being cool using digital media, specifically pervasive and ubiquitous computing devices. Those devices are characterized by being small enough to be carried around as other small objects that surround children (e.g., toys, phones, clothes, pencils, books, stones), or they may be embedded in the environment like other resources we know—in playgrounds, computers, blackboards, and furniture. In essence, these new digital artefacts and environments should be designed in a way that allows them to become a natural part of playful activities.

Compared to traditional artefacts, these devices present new opportunities. They can process information, sense the environment via sensors, interact with it physically through actuators, and communicate with other devices through wireless networks (for early thoughts on ubiquitous gaming in the context of construction kits, see Sargent, Resnick, Martin, and Silverman’s, 1996, list of “Twenty Things to Do With a Programmable Brick”). Apart from actually making the technology work, one of the biggest challenges is to find out how activities involving children, artefacts, hardware, and software can be brought to play in a way that feels natural and fun. The digital media should not be a separate distraction from the play, not in the sense that it should be invisible, but rather that it should be an integrated part of the activities. In order to test our ideas, we have begun designing playful activities for children involving pervasive computing.

StarCatcher: A Simple Example

In this section, we describe an existing pervasive game, StarCatcher (Brynskov et al., 2005). In the next section, we use our framework to elaborate on the game by introducing features that are more complex and, hopefully, more entertaining.

StarCatcher is a simple version of Capture the Flag (Figure 6). It can be played by two or three teams in any urban area of approximately 500x500 meters, and the objective is simply to catch a virtual star (Figure 6b). The first team to reach its location and run to its home base wins and become Star Winners, and the game is over. The losers get the Loser Star (Figure 6c). There is also a slightly more complex version: If the team with the star is intercepted by the other team before it reaches the base, it drops the star, and both teams must visit the home base before it can be picked up again.

Technically, each team is equipped with a mobile phone and a GPS unit (Figure 6a). The game software runs as a client on each phone and on a
central server. The clients send their positions to the server, and the server returns the game state. The phone interface is a map of the close proximity of the team (50-m radius) in which each team is represented by a pulsating dot.

The framework described in section “Activities” offers two ways to describe the game: either focusing on (a) relationships between actions or (b) on the habitats. Figures 7 and 8 use the formalism introduced in Activities to diagram the activity Playing StarCatcher. Figure 7 shows the simple version, Figure 8 the elaborated one. The + and - marks represent the desires of the participants: both team A and B strongly desire to win.

As appears from the diagrams, our game specification uses half-baked actions, as argued in The Habitat Concept. For example, in Figure 8, there are three actions with the same verb: “B walks,” “B walks home,” and “B walks with S.” In the game, these variants have quite different consequences and, therefore, are treated as three different actions. If we magnify these actions, they will look as shown in Table 1, but a description at this level of detail will clutter the diagram and make it useless.

Figure 9 shows StarCatcher described as a habitat. The dashed fat rectangle represents a fictive habitat (see section “Habitats” and Andersen, 2004a). The child C can see its surroundings in the actual (the dashed circle) as well as the fictive world (the dashed rectangle). The reference from the actual to the fictive world is relative (section “Habitats”), since the reference area moves as the access area (the child and the mobile phone) moves. Figure 9 is rather abstract, but in the implementation phase, it will be replaced by a map that shows the exact shape of the actual and fictive habitat and the exact shape of the access and reference area. In the figure, we have indicated that the reference area is different in the actual and the
Figure 8. Diagram of the activity Playing StarCatcher (extended version)

Table 1. The action “walks” with necessary (agent and action) and possible roles filled (goal and sociative); the action has no glue, since it is not a role filled by a participant

<table>
<thead>
<tr>
<th>Role</th>
<th>Agent</th>
<th>Action</th>
<th>Goal</th>
<th>Sociative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glue</td>
<td>Desire (to win)</td>
<td>-</td>
<td>Ability (to be reached)</td>
<td>Ability (to be picked up/carried)</td>
</tr>
<tr>
<td>Filler</td>
<td>B</td>
<td>walks</td>
<td>[home]</td>
<td>[with star]</td>
</tr>
</tbody>
</table>

Figure 9. The habitats of StarCatcher; C = child. A = Agent role, O = Object
The fictive world. In the actual world, the child can see things within a circle centered in the child; in the fictive world, the access area is rectangular due to the shape of the display.

The fictive habitat contains the avatars of the children and the star. The children, therefore, lead a double existence in the real and the fictive world, while the star and the winning position only exist in the fictive world. They are not marked in the actual world. The same difference is found in the actions: catching the star, walking with the star, and winning are fictive, while meeting the other team and walking around in the city are done in both worlds.

The fictive habitat contains two subhabitats: the area around the star where catching is possible and the home base where winning is possible. The action of catching requires two roles: the child (Agent) and the star (Object). In Figure 9, the agent role and catching the star is empty (shown by the white color), so the action cannot be executed. Only when the child moves (the fat arrow) into the habitat will it be able to participate. This is a graphical way of representing the process of unification described in section “Habitats.” Habitats, thus, are loaded with activities that are triggered when qualified participants move into them. This, in fact, is one of the reasons for mobility work (Bardram & Bossen, 2005; Brynskov, 2004); in hospitals, for example, many activities are bound to specific locations, and nurses, doctors, and patients spend much time just moving physically from one habitat to another in order to fill empty roles in urgent activities that are executed as soon as all necessary participants are assembled.

The example shows that our framework is indeed able to conceptualize essential aspects of the game; the dependencies between the actions and, in particular, the relation between actions and physical environment and the interplay between the actual habitat and the fictive one.

In the following, we will describe a new game, Bogeyman, which is an extension of StarCatcher. The purpose is to find out whether the framework is a good support for creative design and whether it can provide new and entertaining ideas while still modeling relevant elements of complex pervasive gaming.

THE BOGEYMAN

If all we wanted to do was to model locative games as simple as StarCatcher, there would not be much need for a framework. But with the expanding opportunities of pervasive gaming (due to technological evolution and commoditization of pervasive computing systems), a whole range of complex issues arise that make formalized support for development of such systems more and more difficult. We can design more sophisticated games using the dynamic interplay of aspects at different levels (e.g., physical, technological and social) but at the cost of complexity. Addressing this complexity at appropriate levels is an important motivation for developing our framework.

StarCatcher actually was implemented and tested; Bogeyman is an extension of StarCatcher and has not been implemented yet, although we do present a core algorithm of the play. The purpose is to test our framework as a design support tool; does it help produce interesting ideas for an extended game?

Activities and Habitats

Let us start with the first two desiderata: we want the following:

16. An emergent game in which simple interactions between more participants can create surprising and entertaining situations (section “Play and Games”).
17. A better exploitation of the interplay between the real and fictive habitats (section “Starcatcher”).
Desiderata (16) concerns the activities of the game, whereas (17) relates to its real and fictive habitats.

In StarCatcher, there was only one thing to collect: the star. In Bogeyman, the purpose is for the children to collect as much candy as possible and avoid being eaten by the bogeyman. While the star was purely fictive, candy may be fictive or real or both. The Bogeyman game is also set in the city (i.e., an area not especially built for the game). However, to a certain extent it does incorporate physical aspects of the surroundings (e.g., the real sewers of the city, which the fictive mice can enter and leave). When the mice are in the sewers, they are invisible in the fictive world.

There can be from two to 10 kids (single players or teams). The technical setup is basically as in the simple StarCatcher game (mobile phone + GPS), with the addition of Bluetooth interaction with places and artefacts.

The objective of the game is to collect as much candy as possible. The purpose, however, is broader, defined as (a) having fun together and (b) developing relationships between the players.

We want some kind of recognizable narrative macro structure (e.g., built upon Greimas’ schema) (Greimas, 1966). A Subject desires an Object, is helped by Helpers, and opposed by Antagonists. In the game, the kids (Subjects) run around and try to catch candy (Object of desire) without being caught by the bogeyman (Antagonist). Whenever they see a piece of candy, virtual or not, they approach it cautiously. If the bogeyman is hiding close by, he will catch them, strip them of all their candy, and eat them. When this happens, they must return to the home base and be brought back to life again (“spawned”) before they can enter the game again. The kids cannot see whether the bogeyman is near. So, in order to avoid him, the kids can get help from the dogs (Helpers). We want these macro roles to emerge from the action dependencies constituting the game.

The dogs, however, are not necessarily interested in helping the kids. But if they receive a sausage, they acquire a positive attitude to the children’s goals and thus desire to act in a way that supports the children.

The execution of the rules controlling the fictive participants conforms to (10) above: the Agent of the action must desire or be obligated to act, and all participants must be able to do so. This means that (a) the software must create desires in the software Agents that are understandable to the children, (b) that actions are only enabled when all participants are in the same location. (b) will involve a lot of physical and virtual movements, which is desirable for a pervasive game.

For example, if the dog is given a sausage, it desires to chase the bogeyman away. But if there is no bogeyman present, the role of the chasee is empty, and the chase-action cannot be executed. Only when the bogeyman appears and fills the chasee role can the dogs execute the action they desire. Processes involving obligations and desires are clearly fictive, whereas processes involving ability is both real and fictive, since the fictive locations of the software agents must coincide with the real locations of the children.

In the section “Activities,” (6)-(9), we listed four requirements to cooperative software agents: comprehensibility, accessibility, joint attention, and intermittent participation. Accessibility means that the action potentials and desires of one participant should be accessible to the others. We can do this by letting the interface show which actions are possible in the current habitat (as already illustrated in Figure 5 and Figure 9) and what the fictive effect of the action is. For example, when a child is near a fictive or real sausage, the display shows that the action of taking the sausage is possible and that it may lead to enrolling dogs in the child’s team. Similarly, when a child approaches a fictive dog, the display suggests that it is now possible to give the sausage to the dog.

If a player wants to know more about a non-player character’s (NPC’s) attitude before interacting with it, he or she may take a closer look first (by using the scan mode described next).
The game is designed in levels of increasing complexity. In the beginning, you only have to find candy. No bogeyman. Then come the mice (an Antagonist) and begin eating your candy. Then cats, which are useful because they eat mice. Then, at last, the dogs and the bogeyman. Dogs protect you from him, but if there is a cat nearby, they may run after the cat instead. If you forget to feed them, they may get angry and bite you. The mice will come out of the sewers, which are connected underground.

When players meet, they can fight each other, but they don’t have to do it. Strength is measured by the amount of candy they have. They can freely negotiate offline, if they prefer (e.g., that the weaker player gives some candy to the stronger team) (security payment). They also can have a virtual fight throwing different objects and power-ups at each other. If they fight and one of them loses, the winner takes all the loser’s belongings. If you find other players’ home bases, you can strip them of all their belongings. A base is defined by a Bluetooth tag. You probably have to look around a bit in order to see it. When you reveal their base, their icon explodes, and they must come to you and collect it to set up a new base. They do this by placing their tag somewhere else.

According to (11) in section “Activities,” execution of actions changes the glue binding

---

**Figure 10. Dependencies between actions in Bogeyman**

![Diagram of dependencies between actions in Bogeyman](image-url)
participants to other actions, and so changes of these dependencies must be an integrated and visible part of the game. Figure 10 shows the dependencies between the actions. As in section “Activities,” we abstract the participant-glue-role relations into three main relations: possibility, necessity, and impossibility in order to maintain overview. To give an example, the capture of the children by the bogeyman is prevented by him running away, which is triggered by dogs chasing him or by the children running away. The capture also can be prevented by the dogs catching the bogeyman. A capture will prevent children from taking more candy; but, and this is the snag, collecting candy is the one action that makes it possible for the bogeyman to catch the children. Collecting candy is risky.

As appears from the +/- marks, dogs desire to catch cats and eat sausage, cats to eat mice and fish, children to eat candy, mice to eat candy and to multiply, and bogeymen to catch children. On the other hand, mice hate being eaten by cats, bogeymen to be caught by dogs, and children to be caught by bogeymen. Thus, the diagram contains conflicts that can be used in a narrative: mice oppose being eaten but cats think it is a good idea; bogeymen like catching children, but children are dead against it. There are also implicit conflicts: mice eating candy prevents children from doing it, and vice versa.

A cursory inspection of the diagram shows that it also contains explicit and implicit contradictions: the children’s taking candy enables the bogeyman to catch them, which again prevents their taking it; the children’s giving sausage to dogs prevents dogs from chasing cats (which they desire) but enables them to eat sausage (which they also desire). Mice entering the sewers prevents them from eating candy (which they love to do) but saves them from being eaten by cats (which they don’t like). The contradictions and conflicts are built into the rules in order for them to generate an interesting narrative, but a consequence is that we cannot use standard logic to implement the rules.

Two rules used by the system are not shown in the diagram, because they would clutter it completely: cats eat mice obstructs all actions in which mice participate: mice multiply, mice enter sewer, mice leave sewer, and mice eat candy. Similarly, Bogeyman catches children obstructs all actions involving children: children give sausage to dogs, children give fish to cats, children take candy, children take sausages, children take fish, children run away, and children eat candy.

Note that we have only distributed desires to a few actions: as it stands, dogs only know that they like to catch cats and eat sausage; they have no particular desire to chase and catch bogeymen or to chase cats. In addition, they have no attitude toward bogeymen catching or not catching children. The idea is that such derivative attitudes should be generated as an emergent phenomenon as the game progresses. Enrolling virtual agents as helpers and avoiding defections to the enemy should be part of the gameplay. Thus, Figure 10 only works as a seed for the game.

The second desideratum was a better interplay between the real and fictive habitats. We shall deal with that now.

The phone is the primary artefact in the informational habitat. It has three modes that are

---

Table 2. Modes of the client running on mobile phone

<table>
<thead>
<tr>
<th>Mode</th>
<th>Scale</th>
<th>Use</th>
<th>Interaction</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proximity</td>
<td>1:1</td>
<td>Navigate</td>
<td>Walking</td>
<td>GPS</td>
</tr>
<tr>
<td>Global</td>
<td>1:10</td>
<td>Get overview of entire arena</td>
<td>Cursor</td>
<td>GPS</td>
</tr>
<tr>
<td>Scan</td>
<td>10:1</td>
<td>Inspect place, thing, or person</td>
<td>Menu</td>
<td>Bluetooth</td>
</tr>
</tbody>
</table>

---
related to the physical habitat: proximity, global, and scan (Table 2).

In the proximity mode, the player can see all objects of the real and fictive habitat within the field of vision. In the global view, the player can only see other players and the map of the arena, including the player’s own home base. The scan view offers a closer look at something (e.g., the attitude of an animal or the presence of a fictive sausage), but it also obstructs the other two views. Whenever a player is in the scan view, it is reflected by the player’s representation in the global and proximity view. This is useful information to other players because they now know that there may be something interesting to check out at that location. Players can also allow each other actually to share their proximity or scan view, thus enhancing each other’s access to information.

Figure 11 shows the three informational modes of a habitat. The scan and proximity modes are relative references that change as the player moves. The global mode is an absolute reference (dashed arrows), since the whole playing area is accessible all the time. The scan mode differs from the two other in that the access area is defined by the location of the reference and not by the receiver. The reason is that the referent (the NPC in Figure 11) carries a Bluetooth tag that creates an information cloud (access area) around it with information about it. If the mobile phone in Figure 11 is Bluetooth-enabled, it can receive this information.

The mobile phone in Figure 11 always can see the global layout of the game area; it can see more details in the local area around it; and it can scan the NPC since the phone is inside the access area of the NPC.

Figure 11, however, does not capture the interplay between the actual and the fictive in the game. In order to do that, we need the concept of fictive habitats introduced in section “Habitats.” Parts of the game exist in the real physical habitat; this is true of the sewers and the children and possibly of the candy and sausages, depending on how you set up the game—and possibly a child may wish to play the bogeyman. Some of the actions also exist in the real habitat; for example, the action of walking and running. However, the game rules only exist in the fictive habitat, and all the participants in the game must have a fictive representation as well. Otherwise their actions could not influence the game state.

Of course, we can depict this dichotomy as in Figure 9 by drawing the actual and the fictive habitat side by side. But here, the correspondence between the fictive and the actual positions of the participants does not stand out. In the following, we collapse the two diagrams into one. The fic-
The Semiotics of Smart Appliances and Pervasive Computing

Figure 12. Basic setup with two players, the bogeyman, some candy, and the home bases

![Diagram of basic setup with two players, the bogeyman, some candy, and the home bases.]

Figure 13. Space-time plot of discrepancies between the physical and fictive habitat

![Space-time plot illustrating the effect of delaying dog shit (topmost) and the effect of a shadow on autopilot (bottommost).]

tive diagram is the main one, since it defines the game state. If a fictive participant has a real-world counterpart, we show this by adding a shadow to the fictive entity. Thus, the real world is shown as a platonic projection of the fictive world. We distinguish between fictive (only in the fictive habitat), real (only in the real habitat), and mixed entities (both in the fictive and real habitat).

Figure 12 shows the basic setup with two mixed players (C), a mixed bogeyman (B), three pieces on fictive candy (S), one piece of mixed candy (S), two fictive mice (M), and one fictive dog (D).

In addition to candy and sausages, there are a number of special items, power-ups, the players can encounter: a dog shit, a shadow, and a torch.

The shit is left by dogs, and if you run over it, you slip, and your avatar is slowed down. You have to slow down yourself, because generally, if you are not close to your avatar, it shifts to autopilot and begins moving on its own. Holiday!

If you walk carefully past the shit and do not tread in it, you can pick it up and drop it later at your convenience (e.g., in front of another player). The stickiness is, of course, entirely virtual and only causes the virtual representation of the player to move slowly.

The shadow, inspired by H. C. Andersen’s fairy tale, The Shadow, is a way to disguise the actual location of a player. When it is activated, the player’s avatar will use dead reckoning to move in the same direction, regardless of the actual movement of the player. In this way, a player can sneak up on another without being noticed. The problem is that if your avatar meets any other avatar, it gets scared because its master is not present, it flees in an erratic manner, and the child must catch it.

The space-time plot in Figure 13 illustrates the effect of delaying dog shit (topmost) and the effect of a shadow on autopilot (bottommost).

The torch (Figure 14) is a way of showing a virtual or mixed entity A to another virtual or mixed entity B. The torch can be implemented as a highlight or a cursor. In Figure 14, an arrow is used. The arrow belongs to the fictive world and, therefore, exists in the fictive reference area on a level with fictive mice and dogs. Each user can manipulate his or her torch. Since the torch is a pointing device, it should only be visible in the proximity mode—you only point out things to people close to you, and physical proximity serves to delimit the number of addressees.

The torch is thus a way of establishing the shared focus of attention between participants, as required in section “Activities.” From the perspective of semantic roles, the torch is a way for...
an agent (C) to increase the glue between another agent (B) and a role in an activity that involves the entity in focus (A). If a child points to a sausage in the presence of a dog, it strengthens the desire of the dog to play the role of Agent in the affordance of the sausage (namely, that of being eaten by a dog).

The torch can also be used to move the focus of the dog away from the affordance of cats (being chased) to the affordances of sausages (being eaten) or bogeymen (being chased away). A mean way of using this technique is to reveal the other players to the bogeyman and make him hunt them.

The game ends after a fixed time or whenever it is suspended by the players. Players can enter and leave intermittently.

In section “Starcatcher,” we decided that the ability to participate in activities is created by spatial movement: if the participants are assembled in the same place, they are able to participate in the activity. But, as mentioned in “Activities,” the activity will not execute unless the Agent participant desires or is obligated to do it. How do we provide these desires?

**Designing Intelligent Agents**

In this section, we will discuss how to implement the distribution of desires mentioned in the previous section. Mice and children are clearly competitors: how can we make the mice aware of this and possibly begin oppose the children (e.g., by supporting the bogeyman in catching them)? How should the cats react toward chasing mice: on the one hand, chasing them leads to the desired action of eating them, but it also enables the mice to flee to the sewers, which disables the eating. We would also like to generate attitudes between participants: who sides with whom? Will the mice oppose the cats and side with the children (although they compete for the candy)? Will the dogs assume the role of helpers to the children, even if helping out prevents them from chasing cats? In short, will global narrative roles emerge from the interplay between the actions?

Which methods should we use to make this come about? We realize already that since the story contains contradictions, we cannot use standard logic for these inferences. Classical AI is thus not an option. In order to choose a solution, let us first look at our requirement lists, repeated as follows for convenience:

1. Intelligent technology is a kind of technology that is able to contribute positively to activities whose manner of performance we will intuitively call intelligent.
2. Agents should have intentions to act that are understandable to the others (comprehensibility).
3. Agents should be able to enter and leave the activity (intermittent participation).
4. An activity is executed if all participants are able to fulfill their roles to some degree and the Agent filler is strongly obligated and/or strongly desires to fulfill his or her role.

(1) says that the algorithm must refer to networks of actors, not to the individual actor; (6) says that the rules it uses must be known to the human users.

One interesting possibility is the theory of cognitive dissonance (Eskola, 1973). The theory allows us to calculate the relation z between persons A and C, if we know the relation x between A and B and the relation y between B and C. It (regrettably) works well for international politics, so we take the fight against terrorism as an example: if
George W. Bush (A) opposes (x) Al-Qaeda (B) and Al-Qaeda supports (y) the Taleban (C), then George (A) opposes (z) the Taleban (C) (Figure 15). In some cases, it is also possible to calculate z if we know x and the inverse of y, y'. If George opposes Al-Qaeda and Taleban promotes Al-Qaeda, then George opposes the Taleban (Figure 16).

If the Danish Prime Minister Anders Fogh supports George Bush, and Bush supports the policy of waging war against Iraq, then Fogh, too, supports waging war against Iraq. We can use the last observation for action selection, as required by (10): when Fogh supports waging war against Iraq, this means that he desires to assume the Agent role in war activities (which, in fact, he did).

The rules can be made to cover normal means-end planning: if occupying Iraq (means) promotes (y') a stable and profitable supply of oil (end), which George very supports (x), then George promotes occupation of Iraq (z). Or, if occupation of Iraq, which George supports (x), necessitates (y) that the American arms budget is increased, then George must support the increase, too (z).

We can, in fact, also leave persons out of the equation: if the occupation of Iraq promotes (x) a democratic government in the country, and democratic government in Iraq promotes (y) a more peaceful world, then the occupation (indirectly) promotes a more peaceful world.

In all of these very real examples, the following two simple rules have been used (≡ is logical equivalence):

18. \( z_t = (x_{t-1} \equiv y_{t-1}) \)

19. \( z_t = (x_{t-1} \equiv y'_{t-1}) \)

\( Z \) at time \( t \) is calculated as the equivalence of the \( x \) and \( y \) (or \( y' \)) relations at time \( t-1 \). At the same time, the rules provide explanations for the attitudes: \( z \) holds because \( x \) and \( y \) hold. George opposes the Taleban because they promote Al-Qaeda, and he opposes Al-Qaeda.

There is one problem, however. As we have already seen, we cannot assume that such networks of actors and actions are consistent in good narratives. Therefore, we often will encounter the phenomenon that the same relation will be calculated differently if we use two different triangles that share one side. This is the problem of overwriting. George Bush endorses an occupation that leads to casualties; therefore, George ought to be in favor of the casualties (rightmost triangle in Figure 17). However, he also likes his voters, who are opposed to casualties. Therefore, George ought to be opposed to casualties (leftmost triangle). If the leftmost triangle in Figure 17 wins, George may change his attitude to the occupation in order to avoid dissonance (e.g., begin opposing it). If the rightmost triangle wins, he has to support casualties (which is the strategy of his op-
ponent: the young suicide bombers are celebrated as martyrs). Machines do everything in a fixed order, so if the rightmost triangle is processed last, it always will overwrite the influence from the left triangle. In order to allow George Bush to change his mind with respect to the casualties, we introduce a random factor in the application of the rules. Sometimes one triangle is chosen, sometimes the other, and sometimes both. This has two advantages: (1) all the triangles are allowed to have an effect, and (2) the “thinking” of the software agents is not momentary but proceeds piecemeal, sometimes overlooking logical inferences in the first round.

However, endorsing the casualties is really no option at all for George for deeper ethical reasons than voter-tactical considerations. There are basic ethical attitudes that are stronger than others and should not easily be changed. In Western cultures, the respect for human life is (officially) very high and cannot easily be changed (publicly). Instead, contradictions should be allowed to remain in some cases. We therefore introduce a measure of strength of the attitudes: weak, medium, and strong.

We also make a distinction between relations where the first A-term is a person and those where it is an action. A person can have subjective desires that can be changed, but actions do not desire anything. They can enable or cause another action.

In summary, we make four modifications to the original simple calculations.

20. Relations between actions only are allowed to change from an unknown 0-state.
21. Relations between persons and between persons and actions can change if they are not strong.
22. The strength of the feelings of a person toward an action propagates to the actions or persons that enable or oppose the action.
23. Rules are applied with a probability between 100% and 0%.

We distinguish thus between immutable factual relations between actions, and mutable subjective attitudes toward actions.

The reason for (22) is the following: if mice strongly dislike to be eaten and being caught by cats enables the eating, then mice should strongly dislike being caught. Cats look quite differently upon the subject: if they strongly like to eat mice and catching them enables this, then cats should be strongly in favor of catching mice.

(20–22) give the network a certain degree of stability, because they restrict the possible changes.

The following algorithm uses three simple data structures. **Network** is a list of actors and their type, `<actor or action, type>`, where type can be P = participant and A = action. **Relation** is a two-dimensional list representing the relation between actors and actions: Relation[i,j] means the relation from i to j. The relation can be “+”, i supports j, “0”, they are indifferent, or “-“, i obstructs j.

Finally, **Strength** is another two-dimensional list where Strength[i,j] represents the strength of the relation from i to j. The strength can be “++”, strong, “0” medium, and “-“ weak.

The rules for overwriting are as follows: a non-zero relation can be overwritten if and only if the source of the relation (the Atype below) is an actor and the relation is not strong. If George

![Figure 17. Contradiction of attitudes](image)
has very strong ethical feelings against casualties in Figure 17, then the rightmost triangle cannot overwrite it.

The different triangles are classified according to the type of their endpoints: thus, PAA is a type where an actor, type P, has a relation to an event, type A, that has a relation to another event, type A.

It is not possible to evaluate the usefulness of the individual triangle in isolation. The PAA case corresponds to the proverb, “If you have said A, you must say B as well.” It is not included in the following algorithm, but it was in the beginning, since it sounded at least as sensible as PPP, “my friend’s friend is my friend.” In both cases, there are exceptions. An example is George Bush’s dilemma where A = George, B = occupation, C = casualties. Bush can evade the responsibility for the casualties by claiming that they are an unfortunate and unintended consequence of an otherwise good deed. In the PPP case, one can argue that the friend made a mistake in cultivating a disrespected person and, thus, not extend one’s own friendship to this person.

PAA was deleted from the algorithm, not because of its inherent qualities but because its effect in the global network was detrimental, as we shall see next.

The algorithm is as follows, written in pseudo-code:

```plaintext
RemoveDissonance (network, relations, strength)
repeat for all elements a of the network
  record the type of element a in Atype
  if Atype = "P" and Chance() then
    -- agents like themselves.
    set the relation of a to itself to +
  end
repeat for all elements b of the network
  record the type of element b in Btype
  record the a/b relation in x
  if Atype = "P" and Btype = "P" and the strength of b/a ≠ + and Chance() then
    --- if you like me, I like you.
    set the relation b/a to the relation a/b
  end if
  if x ≠ "0" then
    repeat for all elements c of the network
      if a = c or the strength of a/c = + then next
      repeat
        --- strong relations cannot be changed!!
        record the type of element c in Ctype
        record Atype+Btype+Ctype in theCase
        record relation b/c in y
        record relation c/b in y'
        record relation a/c in z
        if y ≠ 0 and chance() then
          choose between two cases
          case (thecase = "PPP" or thecase = "PPA")
            --- relations where a is an agent
            set relation a/c to equiv(x, y)
            set the strength of z to the strength of x
          end if
          case (thecase = "AAP" or thecase = "AAA")
            --- relations where a is an action
            set relation a/c to equiv(x, y)
          end if
        end if
        if y' ≠ 0 and chance() then
          --- here we know the inverse of y
          if (thecase = "PAP" or thecase = "PPA" or thecase = "PPP" or thecase = "PAA")
            --- relations where a is an agent
            then
              set the relation a/c to equiv(x, y')
              set the strength of z to the strength of x
            end if
          end if
        end if
      end repeat
  end if
end repeat
end RemoveDissonance
```
The algorithm moves the network of attitudes toward a consonant type by removing dissonances. If a consonant version exists, it will arrive to it through a number of iterations.

George’s dilemma can run as follows (with the PAA possibility included): the seed is Figure 18: George promotes his voters and strongly promotes the occupation, the voters are against causalities, and occupation promotes causalities. After three iterations with a rule probability of 50%, the situation is as shown in Figure 19: George still strongly promotes the occupation and, in addition, the casualties, while he obstructs the voters. The voters retaliate and strongly obstruct George as well as the occupation and causalities. At this time, George’s attitudes were changed in the simulation so that he now strongly promotes his voters. After a few iterations, this changes the whole picture to Figure 20: now George strongly obstructs occupation and casualties but has gained the love of his voters.

Thus, change of action potentials is created by a single change of attitude that ripples through the network in a change of cause and effects.

If we run the dissonance algorithm on the StarCatcher, we will see two teams opposing one another and opposing everything the other team does. There are no surprises, and this agrees with the fact that StarCatcher is a simple zero-sum game. Things are quite different in Bogeyman, since it contains conflicts and contradictions. The participants have their own agenda and create alliances to further their own interests. Bogeyman requires strategic thinking of the children.

In the following, we make some comments on initial experiments that use Figure 10 as seed. The algorithm is being elaborated and refined at the moment.

Figure 21 shows the attitudes developed after four iterations.

Two main groups seem to be forming: cats, dogs and children on one side; bogeyman and mice on the other side. However, there are still some conflicts: cats and mice like the bogeyman but are opposed by him.

The simulation recorded a log of the attitude changes, and this log can be used as explanations for the attitudes. As the following examples show, these explanations seem quite understandable:

- *Cats promote children* because both oppose mice. Cats oppose mice because cats want to eat mice, which mice don’t like. Children, on their side, oppose mice, because the children want to take candy, which the mice don’t want them to for the obvious reason that then it is gone when the mice come to eat.
The Semiotics of Smart Appliances and Pervasive Computing

Figure 21. Bogeyman after four iterations

- **Cats promote the bogeyman** because they both oppose mice eating candy. The cats’ reason for this is that they are enemies of the mice, who like eating candy—cats are against actions their enemies like to do. The bogeyman, on the other hand, doesn’t want the mice to take candy, because it hinders him in catching the children—it is the children’s picking up candy that enables the bogeyman to catch them.

- **Dogs oppose the bogeyman** because they don’t want him to catch children, which he likes. The reason for the dogs’ opposition is that they like to eat sausage, and this is prevented if the bogeyman catches the children: if they are caught they cannot feed the dogs.

- **Dogs promote children** because neither of them wants the bogeyman to catch children. The reason for the dogs’ opposition to child capture is that dogs like to eat sausage, which is obstructed by the bogeyman catching the children.

- **Dogs promote mice** because both want the dogs to catch cats. The mice’s reason is very sensibly that dogs catching cats prevents cats from catching mice.

- **Mice oppose children** because children want cats to eat mice, which the mice naturally don’t like. Children, on their part, want cats to eat mice because children want candy, and cats eating mice supports their candy hunt since it puts a stop to mice eating the candy, which prevents the children from getting hold of it.

- **The bogeyman opposes dogs** because he doesn’t want the children to give sausage to them, but the dogs like very much to be fed. The reason why the bogeyman opposes this sausage distribution is that it causes the bogeyman to run away, which he does not want, since running away prevents him from catching the children, which is his goal in life.

- **The bogeyman opposes mice** because he wants children to take candy, and mice hinder this. The bogeyman wants children to take candy, because it enables him to catch them, which he very much wants.

After the four iterations, most agents thus have developed understandable attitudes toward most actions and agents. However, a few of them turned out to be somewhat strange. Here is an example: it turned out that cats want dogs to catch them (!) because they did not want mice to enter sewers. Dogs catching cats prevents this. The reason why cats don’t want mice to enter the sewer is that they want to catch the mice, and this is not possible when the mice are in the sewer. The reason for the error lies in the seed: we forgot to add that cats strongly oppose being caught by dogs.

In earlier versions in which the PAA type was allowed, we got really weird explanations; for example, the bogeyman opposed children giving fish to cats because he wanted to catch children, and this prevented the children from giving fish to cats. The triangle type is PAA that lets a person transfer his or her attitude from an event to its effect, which seems pretty unintelligible in our case; there is no reason why the bogeyman should oppose children giving fish to cats, even if he indirectly prevents it by catching the children.
Evaluation of the Algorithm

The small experiment shows the following:

24. The dissonance algorithm does produce understandable attitude changes and divides the participants into groups that support or help one another.
25. The explanations given for the attitude changes are understandable.
26. The value of the individual rules must be evaluated through their global effects on the network through experiments. It was not possible to predict that the PAA triangle in itself would be detrimental.

Let us review our requirement list from the previous sections to see how the algorithm fares:

1. Does it produce technology that is able to contribute positively to networks of activities whose manner of performance we will intuitively call intelligent? The “intelligence” of the software agents is clearly a network phenomenon since attitudes develop in the context of other agents’ attitudes.
6. Agents should have intentions to act that are understandable to the others (comprehensibility). The results show that the attitude changes are understandable and that the agents can provide reasonable explanations.
9. Agents should be able to enter and leave the activity (intermittent participation). Since the algorithm recursively tries to remove dissonance in the network, addition or deletion of agents will not spoil anything but possibly may make the network move toward other equilibrium states.
10. Does it conform to the rule that an activity is executed if all participants are able to fulfill their roles to some degree and the Agent filler is strongly obligated and/or strongly desires to fulfill his or her role. The algorithm works by distributing desires of the agents toward actions and other agents. It thus produces one of the glue-components that causes agents to act. The other, ability, is produced by physical movements.

A drawback of the method is that the effects of changes to the seed and the algorithm are difficult to predict, but this is the price you pay for designing emergent games. The designer to some degree loses control over the product.

Another problem could be that the software agents are too intellectual in the sense that their motivations are caused by reasoning with many steps, which children may not be able to follow. On the other hand, many adventure games have plots that are more complicated.

Evaluation of the Framework

Did the framework give us good ideas as we hoped? It turned out it was very productive in generating ideas for playing with physical vs. fictive reality. It was also very good at associating activities to space and participants to activities in a simple coherent way. Bogeyman took about an hour to invent. In addition, the framework maintained a foothold in the computational world and gave indications of possible technological solutions. It worked as a boundary object connecting the aesthetic and the technical world.

The small simulation showed a possible way of combining games of emergence and narrative games of progression, a combination that was judged impossible in Juul (2005).

Because of its robustness, it is very easy to add or delete participants from the game, which makes us hope that it can be developed into a toolkit for the children themselves to modify or build games.

The diagramming techniques clearly need to be developed further. Small games are easy to specify, but more complicated games need conventions for abstraction and decomposition.
also should be emphasized that we lack empirical evidence that the games specified in this way are actually fun to play!

**Future Developments**

The simulation does not deal with the balance between reasoning about actions and executing them, although the conditions for executing actions are well-defined (see (10) in the section “Activities”).

The next step, therefore, is to orchestrate the percolation of desires in the network and the execution of these desires. One convention could be that an action is not allowed to influence other actions before it is executed; only when an action is realized may it have effect on other actions. An even stronger restriction is that only participants who have experienced the execution of the action are allowed to change their attitudes because of it. In this way, the network changes would be much easier to follow, and both solutions will reduce the part of the network that the algorithm must search and thereby decrease the complexity of the algorithm, which regrettably is \( \Theta(n^3) \).

However, the two conventions are probably too strong. On the one hand, communication probably should count as a kind of execution: when the Bogeyman tells the mice that he doesn’t like them to eat the candy, this should make the mice reconsider their positive attitudes toward the bogeyman with the same force as if he had actually prevented them from eating candy. Similarly, a promise from the children to give sausage to dogs should make the dogs side with the children as if they had actually given sausages to the dogs. On the other hand, some degree of secret reasoning is a good literary trick—it invites the reader/player to reconstruct the possible reasons for unexpected behavior.

But the whole play is, after all, staged for the benefit of the children. In order for them to create their tactics, it seems a good idea to inform the children of the interactions between the software agents so that they can understand what is going on and make their plans accordingly. Literature and films have the same problem: how do we inform the reader/viewer of several concurrent happenings when we can only tell about one event at a time? The solution is to let the characters meet, tell stories about their adventures, and let the reader/viewer overhear the storytelling. Should we do a similar thing in the game so that when the children meet a dog, it will tell them about the feeling and atmosphere in the canine world?

But then again, maybe the children should be required to seek the information themselves, at least at advanced levels. The ultimate criterion for deciding is still: is it fun?

**RELATED WORK**

In this concluding section, we compare our approach to a number of related approaches in which notions like communicative acts, activity, semantic roles, context, and pervasive games are central:

- **The Language Action Perspective (LAP) Community**: This community is inspired by the works of John Searle and Jürgen Habermas on communicative actions. The approach to artifact design in Winograd & Flores (1986) also has been an inspiration to the community. LAP basically views use of IT as the execution of speech acts mediated by technology. The notion of speech acts is also central to this chapter, but we extend the concepts to also cover noncommunicative material acts and the relation between the two. The LAP community has emphasized the communicative type but to a certain degree has failed to realize that communicative actions are intertwined with material acts in actual work processes. There is, therefore, a need to build an integrated theory that encompasses the two (Goldkuhl,
The bias toward communicative actions, for example, can be seen in the DEMO method described in Dietz (2003). It offers a finely grained typology of communicative acts (request, promise, decline, quit, state, accept) but only distinguishes between two material acts (deliver, sell). Another problematic feature, which is also noticeable in DEMO, stems from Winograd & Flores (1986): conversation patterns are primarily described in temporal terms: one act comes before or after another act, like in a finite state automaton. The problem is that mere temporal sequence does not explain the reason for the communicative pattern, and it is difficult to relate to organizational issues. The notion of glue in this chapter offers a more detailed method of description: actions hang together because one action changes the abilities, rights, and obligations of an actor to participate in other activities. Since qualifications and norms are central to organizational analysis (Liu, 2000), our framework seems easier to relate to organization theory.

**Activity theory**: Activity theory originated from the dialectical materialist psychology developed by Vygotsky and his students in the Soviet Union in the beginning of the 20th century. Activity theory goes beyond the popular human-machine dyad and insists on cultural and technical mediation of human activity. Therefore, the unit of analysis includes technical artefacts and cultural organization, and the focus of activity theory is much wider than what has been the core concern of past HCI research (Bødker & Andersen, 2005). But activity theory seems to have the opposite problem of the LAP community: it emphasizes material activities in which a subject applies a tool to change some object, and only recently has spoken and written discourse begun to figure as mediators in activity theoretical analyses of work, and the effort to explore their role as mediators has been limited (Wells, 2002). The chapter has borrowed the notion of activity from activity theory but has extended its three basic concepts—subject (agent), mediator (instrument), and object (theme)—by means of the theory of semantic roles.

**Actor-network theory (ANT)**: ANT is not a real theory but rather a number of related methodologies and assumptions (Latour, 1999; Law, 1987) focused on empirical ethnographic fieldwork. This chapter obviously is inspired by this tradition with its emphasis on the significance of networks and its (structuralistic) insistence that participants in the network primarily acquire their properties from their position in the network and not from inherent resources. We deviate from the tradition by insisting that it is possible to refine the vague concept of actor into a set of more precisely defined roles in a network. The roles we have chosen are borrowed from the theory of semantic roles, and the reason why we believe these roles to be real is that they have been formalized in natural languages over many thousands of years by means of prepositions and case inflexions. We thus view the roles as the result of an evolutionary process in which practical experience through the millennia has distilled relevant distinctions. Another deviation is that our framework is biased toward design of pervasive technology, which is not the case with ANT.

**Object-role management (ORM)**: ORM (Halpin, 1998) is an alternative to traditional entity-relation modeling and object-oriented modeling. "ORM is a method for designing and querying database models at the conceptual level, where the application is described in terms readily understood by users, rather than being recast in terms of implementation data structures" (Halpin,
An ORM model consists of a set of objects that play roles similar to actions in Digital Habitats. In our understanding of ORM, it resembles a slot-and-filler grammar similar to ours and to Fillmore’s “Frame Semantics” (Schalley, 2004).

ORM focuses on databases, design, and querying, and it has strong formal properties (e.g., formalized conversion to ER/OO/SQL and nesting). It makes clear distinctions between issues that belong to the conceptual level (how the users understand the problem domain) and those that concern the implementation (e.g., the OO concept of inheritance tends to become very abstract from the problem domain perspective). Thus, ORM is well-suited to validate a model together with domain experts that know nothing or little about programming, because the ORM model can be populated with sample data in a way that reflects the domain on the conceptual level. ORM is supposedly superior to OO when it comes to transformations over time, since it does not use attributes (although facts can be collapsed into attributes for easy overview), which allows for model changes without major restructuring. On the other hand, it only handles the static properties of a model, according to Halpin (1996, 66), since it is the most stable part of the model (although various ORM extensions have been proposed for process and event modeling).

Compared to ORM, Digital Habitats is well-suited to the handling of dynamic models, and ORM lacks the integration of physical and, to a large extent, pragmatic aspects. However, ORM is simpler and has a more formal and consistent design than Digital Habitats.

**Unified eventity representation (UER):**

UER (Schalley, 2004) is a quite recent, UML-based attempt to develop a representational framework for verbal semantics [i.e., verbs, not speech] that is formal and intuitive at the same time. This means in effect proposing a framework that is in principle computer processable on the one hand, and yet on the other hand whose representations reflect the wealth and flexibility of natural language in an intuitively plausible way and in accordance with our current knowledge about natural language. (Schalley, 2004, p. 1)

Based on Unified Modeling Language (UML), it combines computation-oriented diagrams with a solid linguistic approach to compositional semantics. It proposes a diagram-based framework that is precise and yet potentially underspecified. In the context of analysis and design of pervasive games, it seems too fine-grained; however, much insight can be gained from the analyses and discussions that are presented in Schalley’s (2004) proposal.

**Context-awareness:** Context-aware systems stand out as a fairly distinct field within computer science. The Context Toolkit (Dey, Salber, & Abowd, 2001) differs from Digital Habitats in its technocentric orientation. It is a set of concepts that provides concepts and standards for things primarily at a technical level. There is a set of higher-level concepts, but these are developed on top of the technical concepts rather than the other way around.

**Pervasive gaming theory:** Walther (2005) offers a set of concepts to classify and analyze pervasive games. They target important distinctions and issues, and they may well be used in a generative manner, but compared to Digital Habitats, they are much more general in nature and do not support detailed analysis and design.
CONCLUSION

Based on the comparisons in the previous section, we can identify six areas in which the present approach seems to present advantages:

- It provides a description of action dependencies that is not a mere temporal sequence but is relatable to organizational concepts like qualifications and norms (as compared to the LAP tradition).
- It provides a better framework for describing communicative as well as material acts plus the way they hang together (as compared to the LAP tradition and activity theory).
- It provides an explicit link between human activities and their spatial context (as compared to the technical literature on context-awareness).
- It has an explicit dynamic model that precisely describes the conditions for executing actions (as compared to the ORM methodology).
- It offers a typology of participant roles, based on linguistic evidence that reduces complexity and, therefore, supports design processes (as compared to Actor Network Theory and pervasive gaming theory).
- By using semantic roles, it encourages a system architecture that is verbalizable (i.e., which automatically or semi-automatically can produce understandable descriptions of themselves) (Andersen, 2004b).
- Technological artifacts are described according to their function in activities (Bødker & Andersen, 2005). Some automatic systems function as Agents since they initiate and control the activity, others function as Instruments that do not do anything by themselves, while still others function as Beneficiaries in which other participants perform activities on their behalf.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their detailed critical remarks and good suggestions. Part of this work has been supported by Center for Interactive Spaces, ISIS Katrinebjerg (Project #122). Thanks to our colleagues in the research group Frameworks for Understanding Software-Based Systems, which has been a great inspiration for this chapter.

REFERENCES


with virtual inhabited 3D worlds (pp. 189-211). London: Springer Publishers.


*This work was previously published in Semiotics and Intelligent Systems Development, edited by R. Gudwin, J. Queiroz, pp. 211-255, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 2.13
Analyzing the Factors Influencing the Successful Design and Uptake of Interactive Systems to Support Social Networks in Urban Neighborhoods

Marcus Foth
Queensland University of Technology, Australia

ABSTRACT

In urban residential environments in Australia and other developed countries, Internet access is on the verge of becoming a ubiquitous utility like gas or electricity. From an urban sociology and community informatics perspective, this article discusses new emerging social formations of urban residents that are based on networked individualism and the potential of Internet-based systems to support them. It proposes that one of the main reasons for the disappearance or nonexistence of urban residential communities is a lack of appropriate opportunities and instruments to encourage and support local interaction in urban neighborhoods. The article challenges the view that a mere reappropriation of applications used to support dispersed virtual communities is adequate to meet the place and proximity-based design requirements that community networks in urban neighborhoods pose. It argues that the key factors influencing the successful design and uptake of interactive systems to support social networks in urban neighborhoods include the swarming social behavior of urban dwellers; the dynamics of their existing communicative ecology; and the serendipitous, voluntary, and place-based quality of interaction between residents on the basis of choice, like-mindedness, mutual interest and support needs. Drawing on an analysis of these factors, the conceptual design framework of a prototype system — the urban tribe incubator — is presented.
INTRODUCTION

The area of technology and human interaction is cross-disciplinary and requires many different academic fields and design practices to work together effectively in order to generate a better understanding of the social context and human factors in technology design, development, and usage. This article focuses on the social communication aspects of this field and hopes to establish a greater awareness of the contribution that community media and communication studies can deliver to the field of human computer interaction. It seeks to build a theoretical foundation for an analysis of two interrelated issues, which are discussed in turn.

First, the importance of place and the continued purpose and relevance of urban neighborhoods are established. New media and networked information and communication technologies have not led to the diminishment of local place and proximity. However, they have given rise to new types of social interaction and to new emerging social formations. Understanding the nature and quality of interaction in these new social formations can inform the successful animation of neighborhood community and sociability.

Second, appropriate opportunities and instruments to encourage and support local interaction in urban neighborhood networks are not limited to technology, but technology can be a key facilitator. Thus, system designers and engineers are crucial allies to social scientists in the search for hybrid methodologies that integrate community development approaches with technology design. The article questions whether it is sufficient to appropriate tools originally designed for dispersed online (that is, virtual) communities in the context of community networks (Schuler, 1996) for urban neighborhoods. Purpose-built tools and instruments are required that afford

(a) interactive linkages between the resident’s communicative ecologies of cyberspace and local place; and
(b) personalized social networking between proximate neighbors of choice. Such an approach would allow the nonvirtual and place-based assets in a resident’s portfolio of sociability to become more attractive. It would establish an opportunity to create and to maintain local social ties and, ultimately, to find out who is living next door and who is personally compatible.

From the discussion of these issues, some of the key factors influencing the successful design and uptake of interactive systems to support social networks in urban neighborhoods are derived. Drawing on an analysis of these factors, the conceptual framework of a prototype system — the urban tribe incubator — is presented.

This article seeks to set up the interdisciplinary conceptual foundation necessary to drive a thorough theoretical and empirical investigation into the interaction of people, place, and technology and the way they function together to facilitate access to the social and cultural life of cities. The purpose of this article is not only to introduce and illustrate the issues at stake and to present a design framework but also to stimulate transfer and exchange of knowledge across academic disciplines and especially to invite discussion and comment from a broader interdisciplinary audience. Supporting efforts to build bridges between the social and engineering sciences is paramount to the field of technology and human interaction, and this article contributes to the development of a dialogue between these disciplines. An interdisciplinary approach that brings together views and expertise from sociology, urban studies, interaction design, and related disciplines will assist with efforts to facilitate urban neighborhood community building, social inclusion, public consultation and debate, fair access to local information and services, urban sustainability, and healthier local economies.
The Internet has found its way into many households of urban dwellers in Australia and other developed countries to the extent that Internet access is on the verge of becoming a ubiquitous utility like gas and electricity. The Internet has advanced to become a communication tool that coexists with other established communication devices such as the telephone, short message service (SMS), new media, and face-to-face interaction. E-mail, instant messaging, online chats, and other online applications are now instrumental in establishing and maintaining social ties with family, friends, coworkers, and other peers, thus creating a private “portfolio of sociability” (Castells, 2001, p. 132).

The Internet has entered people’s everyday life and plays a significant role in the communication pattern of urban residents. The Internet has not substituted but supplemented off-line interaction with online interaction (Fallows, 2004; Wellman & Haythornthwaite, 2002). People still chat on the phone and meet face-to-face. However, the Internet as well as mobile communication devices such as mobile phones, laptops, and personal digital assistants (PDA) allow people to maintain social ties in different ways by taking advantage of new features. The mobile phone introduced place-independent communication, and the emerging third- and fourth-generation (3G, 4G) mobile telephony adds audiovisual telepresence. E-mail and SMS afford asynchronous communication and notification mechanisms. Online chats offer broadcast-style, many-to-many communication, whereas private chat rooms enable users to engage in multiple peer-to-peer dialogues. Instant messaging tools combine the features of online chat rooms with ambient awareness by adding availability or other status information to a user’s nickname (e.g., Jean|busy, Lucy|out to lunch).

However, these tools are used more often to connect with family, friends, coworkers, and peers and less with neighbors. The telephone has evolved into a ubiquitous communication device, but it has not contributed per se to overcoming urban alienation. Sociologists such as Wellman (2001, 2002) and Wellman et al. (2003) describe how people construct their social networks with the help of the telephone and other devices. Wellman argues that, while people become more accustomed with the features these tools offer, the nature of the social ties that people establish and maintain changes from “door-to-door” and “place-to-place” relationships to “person-to-person” and “role-to-role” relationships. He creates a holistic theoretical framework that builds on the dual nature in the interplay between community and the individual. He describes the emerging qualities of this behavior as networked individualism.

Residential areas, such as apartment buildings, townhouse complexes, master-planned community sites, and the residents and tenants of these units form the focal point in this article to examine the interplay between people, place, and technology. The results and findings of this theoretical analysis will help to shed light on some aspects of the community question, especially the continued purpose and relevance of neighborhoods in urban habitation, by investigating the ironic relationship between endemic urban alienation and the widespread use of mobile and ubiquitous communications technology by urban dwellers that allows them to interact with each other (Walmsley, 2000).

Before this technology became ubiquitous and entered the everyday life of city dwellers, predecessors and variations had been designed for or had first become popular in workplace-based environments in order to support communication and collaboration among professionals. This was followed later by their diffusion into everyday life and their reappropriation for social use. The act of reappropriation (e.g., from the professional use of a pager to the social use of SMS) implies that there
are opportunities to design and develop purpose-built systems from the ground up, which, instead of merely trying to make ends meet, take the unique requirements into account of the social- and place-based context in which they are used. Tools to animate and network urban neighborhoods require a consideration and treatment of notions of sociability, place, privacy, and proximity in order to take full advantage of the communicative opportunities that this environment offers its inhabitants and the wider society.

PLACE MATTERS: COMMUNICATION AND INTERACTION IN URBAN NEIGHBORHOODS

Tönnies’ (1887) idea of community as Gemeinschaft implies a well-connected, place-based, collective, village-like community. However, this notion of community represents an overly romanticized image of community and ignores more contemporary forms of community that have been explored by recent sociological studies (Wellman, 2001, 2002). Gemeinschaft might resemble Hobbiton in the Shire described by Tolkien (1966). This communitarian notion (de Tocqueville, 2000; Etzioni, 1995) is still referred to frequently in the community development literature, although the homogeneous, egalitarian, and all-encompassing nature of Gemeinschaft is a utopian ideal that is less and less compatible with contemporary characteristics of community as social networks in today’s network society.

Before the advent of modern information and communication technology, human interaction was limited by the reach of the physical presence of self or the representations of self (e.g., letters and photographs) and available means of transportation. The need to socialize and to communicate was usually satisfied with family members in the same household, with friends and peers nearby, at work, or within the vicinity of the neighborhood people lived in. Human relations were door-to-door or place-to-place (Wellman, 2001). The fact that people residing in the immediate surroundings were known also established a feeling of security, community identity, and a sense of belonging — a feeling that clashes with the experience of living in today’s high-density, compact urban environments.

The invention and introduction of new information and communication technologies into society has usually been accompanied by foresights that predict that people will be less dependent on place and location. To an extent, this is true. The phone was the first major invention to introduce personal telepresence and to allow everybody to communicate in real time with others outside their own physical locality. Instead of being restricted to people within proximity of oneself, the phone enabled long-distance communication to maintain work and social relationships. However, it is unlikely that anyone lifts the telephone handset to introduce themselves to a neighbor nearby that they have not met before.

The Internet affords both synchronous and asynchronous applications that enable communication between one or multiple users, one-to-many or many-to-many broadcasts to a closed group, and public announcements to an open audience. The abstract nature of Internet-mediated communication gave rise to the widespread use of the metaphor cyberspace, which visualizes the emergence of a new spatial dimension.

However, people’s bodies cannot be atomized in the same way that their audiovisual representations can be digitized, mediated, and sent across the world. Thus, people depend and will continue to depend on place and locality and on co-located face-to-face interaction. Bits and bytes travel in the virtual space of flows spanned by the Internet, but humans travel in the physical space of flows that modern transportation affords. Place and proximity continue to matter in every socioeconomic context, because there are no Internet applications that can completely substitute
Analyzing the Factors Influencing the Successful Design and Uptake of Interactive Systems

real-time, co-located, face-to-face interaction. This is evident by rising car and air travel sales (Wellman, 2001), by people commuting to work instead of working from home and by the formation of economic clusters, precincts, and hotspots where industries based along the same value chain co-locate to take advantage of synergy effects. Florida rightly argues that “the economy itself increasingly takes form around real concentrations of people in real places” (Florida, 2003, p. 4). In light of new urbanism (De Villiers, 1997) and master-planned residential community sites (Gleeson, 2004), his statement holds true not just for the economy but for society, in general.

Attempts to bridge distance for the purpose of more than just communication have seen initiatives such as telework and distance education, yet they remain at the edge of mainstream usage and have not replaced face-to-face interaction (Dhanarajan, 2001; Gillespie & Richardson, 2004). To enable economic efficiencies, the goal of Computer Supported Co-operative Work (CSCW) and groupware applications is to supplement not substitute place-based work practices.

Wellman (2002) points out that the dichotomies of physical place and cyberspace or of online and off-line are misleading. Even as the Internet grows exponentially, place-based units such as home, work, and school remain at the core of our understanding of everyday life. The Internet and other information and communication technology add new qualities to the portfolio of communication tools available to us, enriching our communicative ecology and adding on to the variety of media channels at our disposal. We do not rely on the central location of traditional meeting places anymore, such as the marketplace or town square, in order to meet with friends and peers. Instead, we use mobile communications technology that we can carry around (e.g., mobile phone, SMS) or ubiquitous communications technology that we can access anywhere (e.g., wireless networks) not to avoid but to negotiate on-the-fly meeting places and venues anywhere and anytime. Teenagers, for example, use their mobile phones to arrange meeting places on the spot; this could be the local café, the shopping mall, or someone’s home (Satchell, 2003). This emerging behavior introduces challenges to conventional understandings of place and public places and opens up opportunities for residential architecture, town planning, and urban design (Castells, 2004; Florida, 2003; Grosz, 2001; Horan, 2000; Mitchell, 2003; Oldenburg, 2001; Walmsley, 2000).

In a lively online discussion about the continued purpose and relevance of neighborhood communities, one participant (eric_brissette, 2004) illustrates the point that having less exposure to neighbors (as opposed to coworkers or friends) does not mean that it is less likely that there are, in fact, prospective friends living in the neighborhood:

I guess it all depends on where you live. I live in a rural town of about 10,000. Most people say “hello” or “good morning” to you as you pass them on the sidewalk. I can’t say I’ve known all of my neighbors well, but I have at least spoken with them enough to know a bit about who they are. Visiting larger cities like Boston or New York makes me feel weird. Nobody looks you in the eye, and everyone seems constantly pissed off, almost like everyone is scared of everyone else ... yet this all seems perfectly normal to them. ... Chances are good that there are people in your neighborhood that share your [interests] or are at least [compatible] at the personality level who you wouldn’t normally interact with on a daily basis.

In today’s networked society, it is questionable to project the image of the rural village and use it as a best practice urban village model for a city because of inherent differences between both places and their inhabitants. Yet, the specific characteristics of a city can give rise to a different model of urban village that acknowledges the potential opportunities that this particular environment offers its residents. For example,
the simple fact that a city accommodates a larger number of residents could offer the individual greater choice and, thus, a chance to find the right social interaction partners.

However, the motivation for and process of the search itself remains to be examined. Getting to know someone in their role as a neighbor is less likely than getting to know them in their role as a coworker or being the friend of a friend. Neighbors may still be part of a resident’s social portfolio, but the communication devices used to maintain these ties are inherently place-independent and ephemeral: A phone call or an e-mail does not distinguish between close or distant friends. Proximity does matter when it comes to physical encounters and face-to-face meetings. Most frequent social ties, including online interaction, are maintained with people who can easily be reached physically; that is, they usually reside within the same city, the surrounding suburbs, or the same neighborhood (Horrigan, 2001; Horrigan et al., 2001). The majority of phone calls, SMS, and e-mails helps the parties involved to coordinate meetings or social gatherings (e.g., to catch up over coffee in a café nearby).

These ties are based primarily on common friendship, workplace, or interest, and not on shared locality. We may be introduced and subsequently get along well with the friend of a coworker who happens to live on the same street, but it is unlikely that we would have found out about that person without the coworker introducing us first.

Many urban neighborhoods are the result of what town planners and developers call master-planned communities. Traditional conceptual models of community development limit action to tangible places of public interaction such as kindergartens, public schools, parks, libraries, and so forth (Gleeson, 2004). This build-it-they-will-come approach lacks engagement with the findings of recent community development research (Gilchrist, 2004; Pinkett, 2003). It ignores both the human factors involved in urban renewal and sociocultural neighborhood animation as well as the potential that information and communication technology can offer urban residents such things as online community networks and location-based new media (Day & Schuler, 2004; Rheingold, 2002).

Gilchrist points out that “community development involves human horticulture rather than social engineering” (Gilchrist, 2000, p. 269). Social encounters in urban neighborhoods cannot be master planned. They are based on coincidence and serendipity. Neighbors meet through friends of friends, who happen to live close by; they meet when walking the dogs or, in some cases, when a local problem affects multiple residents (Hampton, 2003). However, more often than not, they do not meet at all, and even if they wanted to, there is usually little opportunity beyond serendipity. Our preliminary results indicate that the majority of residents surveyed believe, just like Eric, that chances are good that there are people in their neighborhood who share their interests or with whom they are at least compatible at the personality level, people they normally do not interact with on a daily basis. For those who would like to find out about them and who still believe in good neighborhood relations, the question remains: What can be done to avoid relying on good fortune and fate? How can those who want to, coax luck?

A step toward a more strategic approach to develop urban neighborhoods encompasses online community networks (Schuler, 1996). Community networks are integrated online systems designed for residential communities that, so far, have usually been comprised of communication tools such as mailing lists, discussion boards, and newsletters. Ideally, community networks allow residents to communicate and interact with other users and to take advantage of the proximity to other residents in the neighborhood. Thus, these systems have the potential to build a bridge between virtual public spaces and physical public
Analyzing the Factors Influencing the Successful Design and Uptake of Interactive Systems

places and to foster network social capital and neighborhood identity.

COMMUNITY NETWORKS IN URBAN NEIGHBORHOODS

Arnold (2003) states that “for the ordinary citizen, social interaction is the ‘killer application’ of the Internet” (p. 83). This development has sparked an increased interest among researchers from a range of disciplines to investigate online communication and online communities (Preece, 2000). Yet, the majority of the work undertaken so far in this research field focuses on globally dispersed online (virtual) communities and not on the use of information and communication technology for communities of place (Papadakis, 2004).

There is a small but growing body of literature that reports on the use of information and communication technology for community development in place-based contexts, mostly within the emerging discipline that Gurstein (2000, 2001) terms community informatics. However, most of these accounts investigate communities that are in one way or another deprived (e.g., telecenters or community access centers in rural and remote locations; ICT for development and poverty reduction in developing countries). The transferability of these studies to urban settings is questionable. Urban dwellers may think of themselves as being quite well-off and may lack common disadvantages, such as low income or unemployment. Such instances of deprivation could contribute to shared agony, which ultimately may help to establish a collective need for change (Foth, 2004b) and, thus, a reason to make use of technology for action and change. In its absence, however, alternative motivations to form neighborhood community need to be found.

Today, the value of door-to-door and place-to-place relationships in urban neighborhoods seems to be on the decline. Researchers and practitioners endeavor to counter this trend through community networking; that is, the application of Internet-and Web-based tools in residential environments to introduce and sustain local communication and interaction among neighbors (Day, 2002). Although the term is sometimes used broadly in other contexts of community development and community informatics, the focus in this article is on urban neighborhoods.

A residential community comprises people who live or stay in a geographically demarcated area. Such communities are sometimes also referred to as local communities, physically or geographically based communities, or communities of place. Apart from the fact that members of a residential community share the same location or address, they are not bound necessarily by any other common characteristic such as interest, age group, or occupation. As such, residential communities are not communities or neighborhoods a priori. An apartment complex might consist of residents who do not know each other.

A range of research projects have been undertaken to examine whether online community networks can facilitate the process of establishing neighborhood identity. These projects set out to design and implement online community networks for both large and small residential sites with various aims and with varying degrees of success (Arnold, 2003; Carroll & Rosson, 2003; Cohill & Kavanaugh, 2000; De Cindio et al., 2003; Hampton & Wellman, 2003; Meredyth et al., 2004; Pinkett, 2003).

Reaching a critical mass of users is considered to be one of the key criteria of success (Arnold et al., 2003; Butler, 2001; Patterson & Kavanaugh, 2001) and has been reported as one of the most common stumbling blocks: “If you build it, they will not necessarily come” (Maloney-Krichmar et al., 2002, p. 19). This statement seems to be common sense; nonetheless, it provides the opportunity for a deeper analysis of the reasons and motivations for urban residents to communicate,
interact, and get together with other residents and to participate actively in an urban neighborhood network.

Dunbar (1996) suggests that the size of human social networks is limited for biological and sociological reasons to a value of around 150 nodes. Barabási (2003) and Watts (2003) provide a more far-reaching overview of recent advances in network theory and their impact on business, science, and everyday life. Some ideas are crucial in understanding community networks: They usually increase or decrease in size, that is, social network research and systems design need to find ways to capture their dynamics. Their structure is not random or chaotic but follows preferential attachment (rich get richer) and fitness (fit get richer). In the context of communities of place, Jankowski and his colleagues support this thesis with empirical research by pointing out that “those geographic communities already rich in social capital may become richer thanks to community networks, and those communities poor in social capital may remain poor” (Jankowski et al., 2001, p. 113). Hampton & Wellman (2003) support this notion by stating that “connectivity seems to go to the connected: greater social benefit from the Internet accrues to those already well situated socially” (p. 283). Then, the next questions are, what constitutes richness and fitness in urban social settings, how do residents get rich (and become a hub in their social network), and how can community networks facilitate enrichment in a fair and ethical manner?

The reasons and motivations for participation in dispersed online (virtual) communities provide further insight into the answers to these questions. A person suffering from cancer might prefer the expertise, empathy, and perhaps anonymity available in an international online community of cancer patients. Philatelists will find more like-minded people in an appropriate virtual community of interest, such as a newsgroup or discussion board that is open to any Internet user and that is not restricted to the residents of just one apartment complex or one suburb. The impossibility or impracticability of a face-to-face exchange in a dispersed online community does not usually impact negatively upon the value that participants derive from such online interactions. The large number of active online communities tells its own tale.

The core characteristic of such dispersed online communities is their collective nature; that is, they accumulate participants who share a common interest, profession, or support need with an entity that acts as a collective group with a shared purpose. The tools that are used to support online communities, including mailing lists, newsletters, discussion boards, and so forth, are more closely designed toward a many-to-many broadcast approach instead of a peer-to-peer networking approach. They assume a pre-existing motivation to participate in and use the virtual space. In the case of shared interest, profession, or support need, that may be the case. However, in the case of residents of urban neighborhoods, the only shared attribute is place and colocaiton. Apart from occasions where an item of discussion or a topic of interest relates directly to the shared place that residents coinhabit, most interaction is located within place but not necessarily about place. Thus, place and proximity are insufficient attributes to attract residents to a community network and to sustain it. Furthermore, a reappropriation of the tools used to support online (virtual) communities in the context of urban neighborhood networks opens up further issues, because a community of place is inherently different from a dispersed community of interest. In addition, connectivity per se does not ensure community, and proximity does not ensure neighborliness (Foth, 2003, 2004b).

The unique selling proposition that could give online community networks for urban neighborhoods a competitive advantage over dispersed online communities is proximity. Community networks allow residents to interact online and to take and continue online interaction off-line, in
Analyzing the Factors Influencing the Successful Design and Uptake of Interactive Systems

real life, and face to face with other residents who live in the same location. As such, they can be an effective tool for local community engagement and activism, if the community faces a shared problem or a common enemy that provides the required motivation for residents to come together. Hampton (2003) describes the experience with residents in Netville, who faced the prospect of losing broadband Internet access, which previously had been provided to them free of charge. The issue and the presence of a common enemy; that is, the Internet service provider, unified residents in community activism to advocate for a continuation of the service, and the traffic in the online community network (in the form of an electronic mailing list) increased significantly. The unifying vigor of a common problem or issue can (temporarily) transform a certain number of residents into a residential collective and, thus, sustain an online community network (Foth & Brereton, 2004).

In the absence of a common enemy, a shared purpose or a pre-existing village-like atmosphere, are there other reasons and motivations for social encounters to occur and for the formation of residential networks in urban neighborhoods? Examining existing urban communities may help to answer this question. Watters (2003) describes the emergence of clusters of under-35-year-old urban dwellers mostly in America but also in other parts of the word as urban tribes. They represent a social network, a swarming group of friends who live in the same city and who are all connected with each other through strong and weak ties. The interaction between members of urban tribes is facilitated through the use of mobile phones, e-mail, and face-to-face gatherings. Watters (2003) does not mention the use of neighborhood or similar ICT-supported networks, but his account of the behavior of urban tribes allows one to imagine a new generation of purpose-built interactive community networks for residents in urban neighborhoods.

THE URBAN TRIBE INCUBATOR: NETWORKING SERENDIPITOUS SOCIAL ENCOUNTERS IN URBAN NEIGHBORHOODS

The previous section discussed the conditions under which residents might ultimately engage in neighborhood community networks and, thus, talk to people within their vicinity. In order for these conditions to emerge, competitive tools need to be designed that allow residents to find out who is living around them and that facilitate local communication and interaction that so far has relied on coincidence and serendipity. However, conventional community networks do not necessarily address these needs. They are very delicate, organic entities. They thrive only in favorable circumstances (e.g., similar demographic and professional orientation) with special nurturing (e.g., free Internet access) (Hampton & Wellman, 2003; Kavanaugh et al., 2003), and chances are high that, otherwise, they may fail (Arnold et al., 2003).

The findings of these sociological studies provide essential insights for a new design methodology that can guide the successful development of interactive systems and devices that can stimulate local interaction and animate urban neighborhoods. A prototype system of an urban tribe incubator is currently being developed and tested in three urban residential sites in Australia (Foth, 2004a). Action research (Hearn & Foth, 2005) and participatory design (Churchill et al., 2004; Greenbaum & Kyng, 1991; Schuler & Namioka, 1993) play crucial roles in iteratively constructing and testing a successful prototype. The participation of residents in the design and development is essential to integrate the range of communication channels that they use and to allow residents to take social ownership of the system.

The previous discussion of the factors influencing systems that support social networks in urban neighborhoods gives rise to a set of design
Analyzing the Factors Influencing the Successful Design and Uptake of Interactive Systems

considerations that are being integrated into the design of the urban tribe incubator prototype. These are now discussed in turn.

**Size, Growth, and Critical Mass**

Popular services and functions in conventional community networking systems include electronic newsletters, mailing lists, and discussion boards. In order to keep these systems interesting and appealing, content needs to be generated either by a systems administrator or a delegate but ideally by the community of users itself. Thus, a critical mass of users is required to maintain an ongoing supply of discussion board postings and responses, mailing submissions, and newsletter contributions. It requires residents to invest a reasonable amount of time and effort in order to collectively sustain the system’s viability.

The urban tribe incubator may include such collective, broadcast-style, many-to-many functions, but the core will be a residents’ directory that does not require maintenance on a regular basis, unless details have changed and need to be updated. A resident’s personal profile may comprise information about skills, trade, interests, hobbies, and contact details. The profile becomes the virtual representation of a potential node that invites other residents to link to and from. The system does not require users to use the directory on a regular basis in order to interact with all other users. Rather, the system allows users to opt in and opt out as they please and as a need arises by facilitating personalized networking; that is, voluntarily initiating contact and building social ties with people of their choice. Thus, the directory becomes the catalyst for personalized peer-to-peer social networks to form.

The size and growth of the directory itself is in no linear relation to the size and growth of an individual resident’s social network. The system acknowledges different levels of social richness and fitness, and thus, the point of saturation remains a personal preference. If an individual’s personal limit of social saturation is reached, he or she can opt out. In conventional community networks, for example, users usually cannot control how many people will respond to their posting on a discussion board: It may be none, or it may set off an avalanche of responses. In an instant messenger application, however, users remain in control of the social network with which they engage — their private buddy lists.

**Diversity, Individualism, and Choice**

The urban tribe incubator is not designed to host an online community of a particular interest or support need but rather allows for the diversity of individual residents with different interests and support needs to find each other and to form smaller social clusters. The system presents residents with choice in relation to the number and characteristics of communication partners and modes of interaction. It provides easy and convenient ways for residents to identify “birds of a feather”; that is, to find like-minded people with common interests or support needs.

The system raises awareness among residents of who is living around them in order to facilitate peer-to-peer connections. The resident directory that links to individual profiles allows residents to choose what personal information they publish online or whether to keep certain information private or only available upon request. The goal of a resident directory is not to facilitate residents initiating virtual contact first (although it can be used in this way) but rather to simplify the process of strengthening serendipitous social encounters that happen while “walking the dog.” Without an urban tribe incubator, such informal contacts that have the potential to develop into rich interaction may remain superficial and transitory.

The system does not require residents to keep communication within the system but allows them to move it to other synchronous or asynchronous communication platforms and devices. Having access to an online directory, a resident is able to
maintain contact with a new acquaintance and to integrate this contact into their established usage of existing personal peer-to-peer communication devices, such as instant messenger, e-mail, SMS, and online chat.

**Privacy and Social Control**

In order to safeguard privacy, residents have control over their personal information and the scope of their online engagement. Enhanced local sociability is welcomed by most residents but must not come at the cost of losing security and control of the voluntary and selective nature of one’s social networks. Our preliminary results are encouraging insofar as residents seem to be trusting their (yet personally mostly unknown) neighbors with personal details such as name, phone numbers, e-mail addresses, photo, occupation, interests, hobbies, and so forth. In our survey, the majority of residents indicated that they are willing to share this kind of personal information online with other residents in the building.

Nevertheless, issues of privacy and social control have to be translated into appropriate terms and conditions that govern the usage of the system and the interaction among residents of the building. It is imperative to ensure that residents have the chance to opt in and opt out at any time without missing out on any essential information. Hence, it is worthwhile to consider supplementing official online communication channels with public announcements on neighborhood pinboards in prominent places within the building (e.g., parking lot entry, reception or entrance area, manager’s office door, elevators) in order to provide alternative ways to access community information.

**Network of Networks, Identity, and Sense of Belonging**

The urban tribe incubator may resemble more the networked nature of, for example, an online dating site than the collective nature of, for example, an online discussion board. What may emerge from this process of personalized networking (or online dating) is a complex web of social networks that span the anonymous void of the building complex, a web of urban tribes (Watters, 2003). Social hubs will continue to play a crucial role as their bridging links (Kavanaugh et al., 2003) connect different social networks and establish connectivity in the sense of community and solidarity. Drawing on viral marketing strategies (Godin, 2001; Goldsmith, 2002), the incubator allows individuals to cross-invite and introduce peers to the other networks in which they participate, both inside and outside the neighborhood. The feeling of a neighborhood identity and a sense of belonging can only emerge if bridging social links between members of different urban tribes contributes to the formation of a mesh-work of urban tribes that is “networked to the ‘edge of chaos’” (Gilchrist, 2000, p. 264). In this context, identity and a sense of belonging are not derived from the collective feeling of being co-located in the same place but from the feeling of being connected to a group of friends who are part of a greater group of peers living close by.

**CONCLUSION AND OUTLOOK**

The design considerations presented here will guide the development of the core prototype system. We then envision to extend this core with more sophisticated features that, for example, allow users to produce and to exchange creative content (photos, audio, video) through sociocultural animation (Foth, 2006) and that simplify the tasks of organizing and managing social gatherings, such as calendaring, inviting, RSVPs, synchronizing with SMS and e-mail, and so forth. As well, in this environment, the social aspects of the urban tribe incubator can be combined with managerial features that allow apartment owners to interact with the body corporate and tenants with the
on-site management. In this role, the system can manage rates and rent payments, entry notices, mailings and notifications, personalized information on contractors, and house rules, thus adding further value to the system and encouraging uptake and usage. Cross-platform compatibility is key. As such, the urban tribe incubator is anticipated to be a technical framework that can be accessed not only on the home or office computer but also on mobile and other devices.

The future holds interesting outlooks for platform developments. New urbanism, urban renewal, and the move toward more and more compact cities create opportunities to rethink the communicative paradigm of apartment complexes and vertical real estate as well as the sociological qualities of the office environment in which most social software is accessed. The kitchen is associated with the preparation of food, which is an essential part of one’s social life, as opposed to the office, which is the center of professional life. Hence, modern residential architecture often links the kitchen area with the living room in order to form one seamless space that can be re-purposed for entertainment and leisure. In this context, the much-scorned Internet fridge might see a revival as an integrated local communication hub that combines the functionality of a simple touch-screen display interface, a ubiquitous instant messenger, and a synchronized resident buddy list with location-aware services and groupware functionality. The rationale for choosing the fridge is not based on the inherent cooling functionality of the fridge itself, but its position and prominence within the environment of many urban homes.

The article contributes to substantiating a new zeitgeist of designing residential community networks for urban neighborhoods, which is characterized by combining current understandings of social networks inherent in Wellman’s theory of networked individualism with the affordances of ubiquitous communication devices and applications for personalized place-based networking such as the Internet, instant messengers, and mobile phones. Putnam (2000) argues that “the Internet will not automatically offset the decline in more conventional forms of social capital, but that it has that potential. In fact, it is hard to imagine solving our contemporary civic dilemmas without computer-mediated communication” (p. 180). If online community networks for residential communities are designed to include features that cater to both collective and network interaction, then they have the potential to contribute to the creation of neighborhood identity and to increase network capital and social capital in urban environments (Florida, 2003; Huysman & Wulf, 2004; Quan-Haase et al., 2002; Watters, 2003). Thus, they may prove to be a milestone in the quest to animate urban neighborhoods, to revive forms of civic engagement in society and to enact global connectivity for local action, in order to move from the vision of the global village to a new understanding of the urban village.

REFERENCES


Arnold, M., Gibbs, M. R., & Wright, P. (2003). Intranets and local community: “Yes, an intranet is all very well, but do we still get free beer and a barbeque?” In M. Huysman, E. Wenger, & V. Wulf (Eds.), Proceedings of the First International Conference on Communities and Technologies (pp. 185-204). Amsterdam, NL: Kluwer Academic Publishers.


Analyzing the Factors Influencing the Successful Design and Uptake of Interactive Systems

information and communication technology (pp. 640-645). Hershey, PA: Idea Group Reference.


Analyzing the Factors Influencing the Successful Design and Uptake of Interactive Systems


*This work was previously published in International Journal of Technology and Human Interaction 2(2), edited by B. Stahl, pp. 65-82, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
ABSTRACT

This article describes a study clarifying information systems (IS) designers’ conceptions of human users of IS by drawing on in-depth interviews with 20 designers. The designers’ lived experiences in their work build up a continuum of levels of thought from more limited conceptions to more comprehensive ones reflecting variations of the designers’ situated knowledge related to human-centred design. The resulting forms of thought indicate three different but associated levels in conceptualising users. The separatist form of thought provides designers predominantly with technical perspectives and a capability for objectifying things. The functional form of thought focuses on external task information and task productivity, nevertheless, with the help of positive emotions. The holistic form of thought provides designers with competence of human-centred information systems development (ISD). Furthermore, the author hopes that understanding the IS designers’ tendencies to conceptualise human users facilitates the mutual communication between users and designers.

INTRODUCTION

As information systems (IS) increasingly pervade all aspects of everyday life, of utmost importance is how applications of IS are adjusted to human action. In particular, in current information systems development (ISD) it is essential to take into account human characteristics and behaviour; that is, to humanise IS (Sterling, 1974). In the same vein, Checkland (1981) argues that ISD should be seen as a form of enquiry within which IS designers’ understandings regulate an operationalisation of their intellectual framework into a set of guidelines for investigation that require
particular methods and techniques for building the system. Regarding the humanisation of IS, a notion concerning the nature of the human being is a crucial element of the intellectual framework. As a consequence, within this kind of enquiry, the way humans are taken into account in ISD is dependent on the operationalisation of the IS designers’ conceptualisations of users. With respect to human-centeredness, attention should be paid to the fundamental qualities of people without any explicit or implicit domination of the other elements of IS, such as data, formal models and technical appliances, or managerial belief systems that treat humans instrumentally. This is necessary in order to conceptualise humans in their own right, and thus avoid the reduction of humans to something that exists only in relation to particular instrumental needs and purposes (cf. Buber, 1993).

Of essential importance is the nature of IS designers’ insights into human characteristics and behaviour that are essential with respect to the IS-user relationship. The most crucial insight regarding human-centred design is to be able to conceptualise users as active subjects comprised of physical, cognitive, emotional, social and cultural qualities, an insight which is the prerequisite for design that promotes subsequent user acceptance and satisfaction. Yet conspicuously absent from contemporary IS literature are empirical studies investigating IS designers’ conceptions of the human users, which have been studied more intensively two decades ago when the systems designers’ inadequate view of the user has been stated to be one reason for the behavioural problems often experienced while implementing IS (Bostrom & Heinen, 1977; Dagwell & Weber, 1983). Also, the lack of knowledge of human needs and motivation on the part of the systems designers has been claimed to cause IS implementation failures (Hawgood, Land & Mumford, 1978). Further, Hedberg and Mumford (1975) have defined the nature of the view of human being held by systems designers as an essential factor in the IS design process. The systems designers’ view of the user is also included in some studies as one of the targets of value choices during the ISD process (Kumar & Bjørn-Andersen, 1990; Kumar & Welke, 1984) and is therefore defined as a value rather than an insight in these studies. Dagwell and Weber (1983), in their replication study, rely on Hedberg-Mumford’s definition of the concept but also refer to Kling (1980), “we know very little about the perceptions that computer specialists have of the users they serve and the ways in which they translate these perceptions into concrete designs (p. 47).” Bostrom & Heinen (1977), in turn, define systems designers’ assumptions of people as one of the system designers’ implicit theories or frames of reference. These previous works do not take an explicit stance toward the definition of the concept “conception,” and do not align the nature of conceptions in detail. For instance, from where do conceptions derive their origins, and what is the nature of those conceptions? In a more recent study, Orlikowski and Gash (1994) discuss their definition of the IS designers’ views. They elaborate the concept “frame of reference” by comparing it to the concept “schema” (Neisser, 1976, pp. 9-11), “shared cognitive structures” or “cognitive maps” (Eden, 1992, pp. 261-262), “frames” (Goffman, 1974, pp. 10-11), “interpretative frames” (Bartunek & Moch, 1987, p. 484), “thoughtworlds” (Dougherty, 1992, p. 179), “interpretative schemes” (Giddens, 1984, pp. 29-30), “scripts” (Gioia, 1986, p. 50), “paradigms” (Kuhn, 1970, p. 43), and “mental models” (Argyris & Schön, 1978). They end up by defining their own meaning for the concept frames as a general concept of shared cognitive structures, not especially regarding humans.

This article describes a study which aims to clarify IS designers’ conceptions of users of IS by drawing on in-depth interviews with 20 IS designers. The analytical choices carried out in this study regard IS designers’ conceptions of users as experiences inherent in their lifeworlds, particularly during the different phases of ISD.
The lived experiences build up conceptions that form a structure of meaning, which incorporates a continuum of levels from more limited understandings to more comprehensive notions; that is, different levels of thought reflecting variations of the designers’ situated, practical knowledge. In this way the development of IS is understood as knowledge work. It is an intellectual and personal process which takes its form according to the conceptions of the performers of the process. IS designers are then applying the ISD methodologies according to their own observations and thinking (Avison & Fitzgerald, 1994; Hirschheim et al, 1995; Mathiassen, 1998). Then the most important tool for ISD, and a key resource in contemporary IT companies, is the IS designers’ thought and insight (Nonaka & Takeuchi, 1995; Quinn, 1992). Particularly with respect to the humanisation of IS, designers’ conceptualisations of the human users are seen as knowledge that reflects the designers’ competence in humanising IS. In this way, IS designers’ conceptions may be seen as intellectual capital that mirrors the know-how, practices, and accumulated expertise of practitioners within a particular profession (Kogut & Zander, 1992).

In what follows, first, the assumptions informing this study are presented by introducing the interpretative approach referred to as phenomenography. Second, the resulted forms of thought are presented and discussed. The IS designers’ conceptualisations presenting their mental schemes of the human user result in three forms of thought, revealing both context-centred and human-centred understandings of what humans are. These three levels of understanding indicate that IS designers tend to conceptualise uses in terms of technology, business and work, and that seldom are users taken into account according to their human qualities. Finally, the different levels of conceptualisation are discussed in relation to human-centred ISD.

DIFFERENT LEVELS OF UNDERSTANDING

This study merges with the principles of phenomenography, which is a qualitatively oriented method of empirical research for investigating the different ways in which people experience aspects of reality (Marton, 1981; Marton & Booth, 1997). Essentially, phenomenography is about individual meaning construction, which results in a conception. The primary focus is on the structure of the meaning of conceptions, which are seen in the light of the phenomenological notion according to which person and world are inextricably related through a person’s lived experience of the world (e.g., Husserl, 1995; Merleau-Ponty, 1962). Our intentionality is seen as qualitatively varying foci on the horizon of our life worlds.
Different Levels of Information Systems Designers' Forms of Thought and Potential

While experiencing the world, individuals form conceptions, including qualitative dissimilarities, which are inherent in the intertwined referential and structural aspects of an experience. Different levels are due to the way the structural aspect and the referential aspect merge with each other. Then an experience is specified by the analytical distinctions of a structural aspect and a referential aspect (Figure 1). The structural aspect denotes how a particular phenomenon is both discerned from its environment and how the phenomenon's parts relate to each other as well as to the whole phenomenon. That which surrounds the phenomenon experienced, including its contours, is its external horizon. The parts and their relationships, together with the contours of the phenomenon, are its internal horizon. The referential aspect signifies the meaning of the conception. These two aspects are dialectically intertwined and occur simultaneously within an experience. Thus, people create conceptions with respect to the structural aspect’s external and internal horizons of a phenomenon that are dialectically merged with the referential aspect of that particular phenomenon.

On the one hand, conceptions differ in terms of content, and on the other hand, they differ in terms of the extent of the form that a certain phenomenon is experienced, as a part of that phenomenon, or more as a whole. When detached parts of a phenomenon are the focus of thought instead of relating the parts meaningfully to the whole phenomenon, the meaning of the phenomenon is understood in a way that refers to a more narrow view. Respectively, when the focus of thought is more comprehensive regarding the whole meaning of a phenomenon instead of on separate parts of it or even the surroundings of the phenomenon, the more explanatory conceptions are. Further, the more explanatory power conceptions have, the better they support competent action with respect to the phenomenon in question (Sandberg, 2000). Based on these differences, conceptions form a structure of meaning, which incorporates a continuum of levels from more limited understandings to more comprehensive notions. The qualitative variation in the ways that IS designers conceptualise users reflects their different conceptions of users, and simultaneously forms different levels in the designers’ understandings, reflecting variations of their situated, practical knowledge as forms of thought. These forms of thought, in turn, suggest different levels of competence in humanising IS, because the subjective conceptualisations of IS designers refer to their intention of action (Säljö, 1994).

The interview method is accomplished with respect to phenomenographic principles. To promote multiple interpretations within individual designers, first, opening questions with varying perspectives into the different phases of ISD, such as planning, design, implementation, use and maintenance, was incorporated in the interview framework. Second, to sustain the connection between the designers’ reflection and the actual work within ISD phases while the in-depth interviews aimed at achieving mutual and authentic understanding that proceeded from the interviewees’ expressions. The interviews were taped at the designers’ workplaces to maintain the work practice orientation and to facilitate the expression of the connection between the respondents’ immediate experiences and the subsequent conception. Second, the respondents represent a variety of geographical location, age, gender, educational background and work experience. They came from various different application areas of IS practice, such as e-commerce, tele and media communications, groupware, health care systems, office automation and insurance systems.

The analysis of the interview data was carried out against the phenomenographical theory. Since phenomenography does not offer grounds for defining the content of conceptions within data, a coding paradigm was developed from the data in order to facilitate the identification and categorisation of meanings in the data (Glaser & Strauss, 1967; Tesch, 1990; Strauss & Corbin, 1990). The development of the coding paradigm was based
Different Levels of Information Systems Designers' Forms of Thought and Potential

on the phenomenographical principle regarding the intentionality of conceptions. Conceptions are context-dependent and every experience is described in content-loaded terminology (Säljö, 1994.). That is, the descriptions are carried out in terms of the nature of the situational experiences in question. Consequently, the meanings are to be found in accordance with the underlying assumptions concerning the intentional nature of ISD. In addition, since it is assumed in phenomenography that the meanings of the respondents’ mental acts exist in the data and are constitutive of the data (Walsh, 1994), the way meanings are understood in this study should also be in accordance with the types of intentionality existing in the data. Thus, the coding paradigm was finalised by bracketing away any preconceived ideas of what the IS designers’ views might be like (Francis, 1993), and letting the analysis be informed by the underlying assumptions of intentionality in ISD.

First, the data includes utterances that describe various actions and objectives concerning ISD. These expressions indicate intentionality as defined by Hirschheim et al. (1995, p. 16). They state that “IS development is intentional, to the extent it reflects a planned change. It is based on developers’ intentions to change object systems towards desirable ends”, and go on to say that (1995, p. 17) “intentions in systems development are expressed by objectives. These are related to general value-orientations and represent what ‘one ought to do’ or ‘what is good’.” From this it can be concluded, in the first place, that intentionality in ISD is expressed by intentional action. That is to say, IS designers’ descriptions of the actions and means they are involved with when developing an IS reveal the meanings they give to the phenomena they deal with concerning ISD. This notion is in accordance with the principle of contextuality in phenomenography, which denotes that people’s conceptualisations are not detachable, either from their context or the content of the task at hand. This stance also reinforces the interpretative nature of phenomenographical analysis in that the researcher must see the designers’ action as inherently meaningful (cf. Schwandt, 2000). In the second place, as Hirschheim et al. (1995) point out, intentions are expressed by objectives of ISD. Consequently, it is an appropriate way to define that the way the IS designers understand the human user of an IS is revealed through descriptions in which the respondents’ focus of reflection is on the objectives of ISD. That is to say, in addition to the actions and means the designers refer to, the IS designers’ intentions to change object systems toward desirable ends reveal the meanings they give to the phenomena they deal with concerning ISD. These desirable ends or objectives represent the things that are regarded most important in ISD. In this way, the IS designers’ descriptions of action, means and objectives also implicitly indicate value orientations included in the process of ISD. Therefore, the described actions and objectives represent the things that are regarded important, and thus reveal the referential aspect in terms of intentionality as an implied value orientation. This means that the initial referential aspects of conceptions may be found in utterances in which the designers refer to their way and means of building systems and the objectives of their actions.

Second, the data includes descriptions in which the respondents’ thoughts are attached to human objects. These descriptions of people indicated human features and also value orientations toward people. Often these descriptions also included expressions which indicated emotionally toned reactions. These kinds of expressions within the data indicate intentionality that is in accordance with Uljens (1991), who states that the process of qualitative individuation of a mental act has been done when an object and a psychological mode, referred to as an attitude, is shown. In other words, how a particular object of thought is experienced denotes the respondents’ attitudes toward the phenomenon that is being reflected on. In brief, the inherent meaning of an utterance may be seen as
the correlation between the what- and how-aspects in that they are not detachable from each other, but are interrelated in a particular logical way, indicating what a particular phenomenon is, in what it is revealed, and what kind of values and attitudes are related to it. As described above, the search for the meanings in the data, data analysis, was initiated by establishing a coding paradigm, which suggests that the meanings in the data are found in utterances in which the designers refer to their actions, means, and objectives concerning ISD, as well as to human characteristics.

The subsequent analysis procedures followed the idea, firstly, of iterating between the meaning (referential aspect) of single statements, their surrounding statements, and the data as a whole, and second, iterating between the features that reveal different levels in these meanings (structural aspect). The analysis appreciates the phenomenological notion of “Lebenswelt” in that the continuum of levels from more limited forms of thought to more comprehensive understandings reflects associative connections within the designers’ conceptions rather than direct, law-like logical relations between them (cf. Husserl, 1995). The analysis revealed the designers’ conceptions as different levels of understanding, which appear as the separatist, functional, and holistic forms of thought. These forms of thought reveal three different levels of understandings incorporating both context-centred and human-centred notions. In the descriptions associated with the context-centred conceptions, the designers’ focus of reflection is on technology, work, and business. The human-centred conceptions deal with knowledge, emotions, and the designers’ selves. The following description delineates the specific contents of the conceptions as parts of the three forms of thought by highlighting some exemplary conceptions.

**THE SEPARATIST FORM OF THOUGHT**

The most partial way in which the IS designers conceptualise humans is through the separatist form of thought. It demonstrates how IS designers see humans within the affordances and constraints of contemporary IS and their development as separated from fluid and coherent interactions. Within this form of thought, the user is positioned outside the IS designers’ awareness through objectivist conceptualisations. An objectivist pattern is evident in the context-centred separatist conceptions, which reflect understandings according to which reality exists independent of humans and can thus be understood independent of humans (cf. Lakoff, 1987; Orlikowski & Baroudi, 1991). In this case, when discussing human-centred issues of ISD, the designers’ focus of reflection is directed to technology, job titles, and market mechanisms. For example, within the conception of “the human being displaced by technology” IS designers refer to humans in terms of technology:

R: What are these requirements and wishes like? Could you tell me more about them?
D16: Well, because it is a question of – let’s say — a feedback channel that our company offers as a product to its clients, it means that if the client purchases, for instance, a datanet-based customer network, they have datanet and router accesses through which they operate between their networks and use the whole telecommunication network. Then there are a lot of this kind of usability issues, response times and load percentages, or in a way, how it (telecommunication network) sort of behaves, what happens there.

In the above interview extract, the designer considers the customers’ needs as a piece of software — “a feedback channel” — and the main point that emerges is how this item of software works with the functions of a telecommunications network. The designer’s train of thought becomes...
focussed on technology instead of human-centred issues and needs, such as how the software is built in regard to the humans that will be using it. In the same vein, within the conception of “the human being as a market” the designers make use of expressions which show their intention is to build products that are profitable and, therefore easy to sell. Yet they do not base their intentions upon human features, such as spontaneous and mood-related online behaviour that could be a prerequisite for selling their products (e.g., Hoffman & Novak, 1996; Bellman et al, 1999):

**D5:** It is more reasonable to develop a mass product which has a lot of users. The point here is that then it can be copied and sold.

In contrast to notions that emphasise understanding human consumption behaviour, the IS designers adhere to the idea that humans are simply a featureless mass of consumers who form a market for IT products. Because the above conception does not incorporate any human characteristics, but refers to a mass market, it is thereby making a clear distinction between the market and the features of the people assumed to form that market. For this reason, the conception appears as objectivist. This conception also implies a predisposition according to which the current development of IS as an industry is that of a rational institution which produces mass culture by reducing humans to members of a mass (cf. Slater, 1997). Furthermore, humans become separated attitudinally from IS and their development due to a presumed lack of technological knowledge, and thus are forced to encounter disparaging attitudes. The separatist human-centred conception of “the technology-illiterate human being” produces accounts according to which the most distinct characteristic of humans is that they are ignorant of technology, specifically computers, software and ISD methodologies. In particular, this illiteracy is seen as a contrast to the literacy of the IS designers:

**R:** Have you ever wondered why people behave in that way — that they cannot say what they want from the system?

**D17:** I think that it’s because they don’t know how these [IS] are defined. If one doesn’t know these methods, one can’t do it. That is the biggest reason, not that they aren’t willing to say what they want but they don’t have the know-how.

Beath and Orlikowski (1994) report similar findings in their analysis of a relatively new representative of the ISD methodologies’ rationalist tradition, information engineering (IE). According to the analysis, the IE text creates and sustains both implicitly and explicitly a dichotomy between users and IS designers by characterising the users as technologically ignorant in regard to the use of technology. When operationalised, these characterisations are likely to generate nonviable and unsatisfactory interactions between users and IS designers. It seems also that the designers do not consider the weaknesses in users’ knowledge and thought as an issue that deserves to be taken into account in design. However, when humans are included in the design considerations, the weaknesses in people’s thinking should be understood as natural flaws in human cognitive behaviour that can be appropriately guided, or even prevented by adequate design (Kirs. Pflughoeft, & Kroecck, 2001; Norman, 1989; Robillard, 1999). The idea then is that designers should have awareness of and be able to recognise these erroneous tendencies in users in order to carry out IS planning and design with the express aim of preventing people from committing faulty actions during computer use, rather than conceptualising users as technologically ignorant.

Moreover, negative emotions and physical stress symptoms have the effect of separating humans from IS. The conception of “the computer anguish human being” reveals views acknowledging that IS cause negative emotional arousal in users. These reactions are manifested as negative attitudes, resistance, fear and discomfort in
situations where people are confronted by plans for the future use of computers or in situations in which individuals are using computers:

**R:** How in your mind do people learn to use software?
**D6:** ... I have also met users who have so much fear of the user interface that they don't dare to explore or try anything, they just do what is familiar and safe.

These conceptualisations are consistent with statements concerning the widespread existence of technophobia (Brosnan, 1998a). Besides being an obviously unpleasant and undesired experience, negative emotions, such as anxiety and fear, make people’s behaviour withdrawn and elusive by narrowing their action (Fredrickson & Branigan, 2001), as well as decreasing the quality of their performance (Brosnan, 1998b).

In brief, within the separatist form of thought, the human being becomes separated from viable interactions with both the IS designers themselves and IS. This is due to a tendency to an objectivist conceptualisation, which blurs the designers’ thought to such an extent that people are no longer recognised as humans. The overall narrative style of this form of thought is reminiscent of the style of a nomothetic science reflecting technical, strategic views aiming at controlling the IS-related social system with the technical system (cf. Deetz, 1996). Further, people become separated from the development of IS due to disparaging attitudes inherent in designers’ assumptions that users are technologically ignorant. Moreover, this form of thought brings to the fore human characteristics, such as negative emotions, which are seen as an obstacle to a viable IS-user relationship.

**THE FUNCTIONAL FORM OF THOUGHT**

The functional form of thought consists of conceptualisations in which humans are seen to act in an insubstantial manner, adapting to the external functions of technology, work tasks and the way the IS designers themselves use computers. Within this adaptive response, positive emotions are required in order to create and sustain viable interactions with IS. In this way, the IS-user relationship is seen as functional: the action of people is seen as determined by their external environment, and the role of human emotion is to facilitate this process. The different conceptions that build up this form of thought reveal a behaviourist understanding of the human being.

For instance, the conception of “the invisible human being” denotes humans as using IS in an insubstantial manner. Typical of this conception is the belief that there is a user who uses an IS. Yet the user is not characterised further but is assumed just to use the system:

**R:** If you think of a situation where you are creating an application, who do you think you’re making it to?
**D16:** Hm....

**R:** Do you think of certain types of people or how does it show that you are making it for people?
**D16:** I don’t think of particular types of people but I think that the human being is in some sense always a part of the system. If it is a system that has a user interface so there must be somebody who uses it. Even if it is a system that runs by timer initiation, there must be a user interface, too, for setting the timer parameters in the system, so there must be somebody to use it, too. To my mind there is always someone using the systems, they (systems) are not fully automated.

A functioning relation between people and IS is thus acknowledged, but this does not include any features originating from the mental, social
Different Levels of Information Systems Designers’ Forms of Thought and Potential

or cultural human modes of being. In other words, humans and their behaviour are understood as purely physical-organic responses to technology, as established in the tenets of Skinnerian behaviourism (Skinner, 1938, 1991). Similarly, the IS designers’ conception of “the human being behind the process of work” denotes humans in conformity with behaviourist thinking. Within this conception individuals are seen in terms of their work tasks or organisational work processes. Characteristic of these conceptualisations is that the people performing the tasks are not portrayed further, but are assumed merely to use IS according to the external task flows:

R: How would you define users’ needs?
D8: They consist of the utilising organisation’s needs at all levels, beginning with what the people need in order to continually do their work tasks, and ending with the things that the organisation expects from the system, what can be abstracted from the process and be used to develop and control action.

Here, human action is seen as a series of direct responses to external work tasks issued to people. Zuboff’s (1988) well-known distinction between “automating work” and “informating work” highlights the difference between implied behaviourist and nonbehaviourist assumptions concerning human action in computerised work (pp. 9-10). Automating work refers to deploying technology in ways that increase the self-acting and self-regulating capacities of technical systems, which are expected to minimise human intervention. Because human intervention is minimised and machines perform the work tasks, interactions between individuals and computers become determined by the structure and sequence of computerised workflows to which, in turn, humans are supposed to respond. Zuboff’s term of automating work, thus, implies a behaviourist assumption of humans and their behaviour.

Quite the opposite is suggested by the term informing work, which adds to the automating view of work in that information technology can be used to automate, but at the same time, it has the ability to translate the automated activities into a form that renders work processes, objects, events and behaviours visible, knowable and sharable for people (Zuboff, 1988). That is to say, within the interaction of humans and computers, people actively observe, interpret and share the information which is mediated to them by IS. They do not just respond like marionettes to the information offered by IS, but actively construct their own conceptions of the computer-mediated tasks they are given and act according to their own interpretations of the particular situation. Thus, in order to accomplish fluid and coherent interaction designs between humans and computers in regard to particular tasks, the users’ mental models, especially those concerning the tasks submitted to them, should also be designed (Norman, 1989; Preece, 1994).

Also typical of the functionalist form of thought is that the role of human emotion is to facilitate people’s adaptation to technology. Within the conception of “The techno-enthusiast human being” the designers depict positive emotions, such as enthusiasm, as essential features in humans:

R: Do you think there are common features in those people for whom you have built systems?
D17: Well, at least during very recent years, it has been enthusiasm.

In particular, positive emotional reactions in people are seen to be induced by technology. Positive feelings are especially seen as a prerequisite for the successful use of IS. These conceptualisations reveal a functional understanding of positive emotions. Whereas negative emotions are associated with specific tendencies, such as an urge to escape or to avoid disquieting things, positive emotions seem to spark changes
in cognitive activity in addition to producing behavioural tendencies (Fredrickson & Branigan, 2001). Therefore, the IS designers’ accounts of positive emotions as a prerequisite for the use of computers imply an understanding of the role of human emotional features in promoting successful functioning.

To sum up, within the functional form of thought humans and their behaviour are understood from a behaviourist stance, which renders human substance only as physical and organic by nature, denoting that the movements of people can be explained by the laws of mechanics (Wilienius, 1978). However, this form of thought adds to the previous separatist way of thinking in so far as humans are actually depicted as performing tasks with computers, whereas in the separatist form of thought the conceptualisations either totally omit human features or humans are seen as unable to use computers. In addition, the human emotional feature that is recognised in this form of thought appears as positive — even though functional — in nature. This way of thinking acknowledges humans as users, and therefore is more developed than the previous separatist form of thought.

THE HOLISTIC FORM OF THOUGHT

The most developed form of thought by which the IS designers conceptualise humans as users of IS is the one characterised as holistic. Its holistic quality is revealed in several ways. First, unlike the preceding forms of thought, the designers recognise a number of human characteristics. Second, these observed human features are often seen to coexist or intertwine with each other. Third, these conceptualisations suggest that the relationship between users and designers, as well as the IS-user relation, is a reciprocal process, including characteristics typical of human behaviour.

To begin with, the conception of “the human being reflected in technology” reveals the specific goal of constructing computer interfaces with human-like features: the interaction between people and computers is then envisaged as enriched with dialogues conveying both the rational and emotional meanings of the information in question (e.g., Nakazawa, Mukai, Watanuki & Miyoshi, 2001). Respectively, the depictions of various human features in technology reveal understandings suggesting human features built into technology render the interaction between users and IS as resembling the interplay of cognitive, emotional and social aspects that occur between humans:

R: What kind of user interface do you think that people would want to use?
D4: I strongly believe that 3D interfaces are coming. They could offer kind of human-like facial features as agents, which would bring a human sense to the systems. The third dimension could also be utilised so that interfaces become tangible and accessible.

Further, the context-centred conception of “the human being as an organisational learner,” which highlights people as organisations which learn about their own work processes, refers indirectly to learning, which stresses both cognitive and social human features. Collective cognitive features are referred to as an organisation’s ability to form new insights into its work processes and to guide the deployment of IS effectively (Robey, Boudreau & Rose, 2000). A social dimension is also implied when it is assumed that people learn as an organisation:

D8: Needs are prone to change rapidly, especially after the implementation of the system, because they teach an organisation a lot about itself, and an organisation’s self-knowledge increases and usually needs change in a more clever direction. Then there very quickly happens a sort of ‘learning leap’, which is often experienced as if the system is not valid at all although it is a question
of the organisation’s increased knowledge of its own activity.

Within the conception of “the knowledge sharing human being” the designers open up their view of learning by specifying mutual understanding between users and designers as essential. “It is important to be able to explain things so that we understand each other.” The capability of taking another’s perspectives into account form the core of this conception, which highlights knowledge sharing as a particularly important instance within the processes of organisational learning. Knowledge sharing is the link between individual and group learning, and signifies the expansion of individuals’ cognitive maps into shared understandings (Crossan, Lane & White, 1999). In particular, the ability to take the perspective of others into account is an indispensable prerequisite for knowledge sharing (Boland & Tenkasi, 1995). Buber (1993) ascertains that, in order to be able to take others’ perspectives into account fully, one has to treat others as equal human beings and respect the current circumstances of others. In these kinds of relationships positive emotional features, such as care and joy, need to be acknowledged and combined with cognitive and social abilities (Fredrickson & Branigan, 2001).

Moreover, the conception of “the emotionally coping human being” refers to an ability to regulate in a successful way both negative and positive subjective feelings in computerised situations. In this way, the designers see emotional coping in the light of positive outcomes (cf. Folkman & Moskowitz, 2000, pp.648-649):

D8: ... a skilful user always has such peace of mind and attitude. She or he kind of has a better tolerance for stress, and an ability to cope with contradictions in a better way than others. For some reason this kind of attitude leads to a particular resourcefulness and an ability to utilise the system in a more natural way, compared to a person who has some negative emotional features, fear or hostility towards the system, and who then ends up having difficulties with the system due to her/his heavy attitude.

A cognitive aspect is seen as inherent in emotional coping in that it requires that individuals’ recognise their different emotional experiences. However, in addition to these internal cognitive-affective features, emotion regulation refers to the external social and cultural factors that redirect, control, and shape emotional arousal in such a way that an individual is able to act adaptively in emotionally activating situations (Pulkkinen, 1996). While ISD is often seen as a stressful process which requires an ability to endure changing emotional experiences, such as interest and frustration (Newman & Noble, 1990) in recurrent situations of failure and subsequent success (Robey & Newman, 1996), it is understandable that the designers regard people who are able to regulate their emotions successfully as skilful.

Briefly, the holistic form of thought is comprised of conceptualisations that regard humans as cognitive, emotional, social and cultural creatures. The conceptions belonging to this form of thought embody similar basic human modes of being, as shown above. However, the aforementioned basic modes of being emerge in these conceptions as different behavioural affordances. The cognitive mode of being is seen as intellect, reasoning, learning, reflection, understanding and awareness of something. Similarly, the emotional mode of being is conceptualised as empathy, stress, tranquillity, commitment, contentment and a feeling of mastery. Further, the social mode of being is referred to as a need for communication, group learning, interpersonal power and connection, as well as knowledge sharing. These behavioural affordances are seen as incorporated in technology, appearing between humans, or within the interaction of humans and IS.
THE FORMS OF THOUGHT IN ISD

The IS designers’ forms of thought revealed in the results of this study are regarded as important tools for ISD, and are seen to have implications for the ways that humans are taken into account as users within the different situations of ISD. These different situations refer to the phases of ISD such as planning, design, implementation, use and maintenance. The phases are cyclical and intertwining (e.g., Beynon-Davies, Carne, Mackay, & Tudhope, 1999). Planning refers to initiation and requirements analysis actions, including client contacts and definition of user requirements. During this phase the greatest degree of interaction occurs between users and designers (Newman & Noble, 1990). In order to accomplish requirements analysis, the designers should understand many human issues in addition to technical ones (Holtzblatt & Beyer, 1995). Design denotes procedures where requirements are refined and turned into specifications and finally software. Then technical reliability and maintainability of the system, user interface’s applicability for the intended purpose of the system, as well as the aesthetical appearance of the system, are designed (Smith, 1997). Winograd (1995) emphasises that, in addition to technical requirements, the properties of a user interface should meet with the social, cognitive and aesthetic needs of people. Especially within new ubiquitous technological environments, the design of IS-user relationship should focus, in addition to social and cultural features, on individuals’ perceptual, cognitive and emotional space (Stephanidis, 2001).

How would the designers then perform according to their forms of thought? The strength of the designers utilising a separatist form of thought would be technical knowledge, especially the ability to fluently conceptualise issues of design in accordance with objective definitions, a skill that is needed in creating formal specifications. However, the validity of objectifying design issues is dependent on the focus of such definitions. From a human-centred perspective, valid definitions would require being theoretically sensitive to human activity and deriving second-order conceptions from that activity (see Walsham, 1995), rather than creating objectivist conceptualisations, which overlook humans and their behaviour. An obvious disutility would be a tendency to treat users as technologically ignorant, which implies incompetence in social relationships with users.

The designers embracing the functional form of thought possess technical knowledge, and value such knowledge in users. They tend to focus on formal job descriptions, external work tasks and individuals’ task productivity. A deficit from a human-centred perspective would be the tendency to overlook human issues and to focus instead on the functional purposes of IS; that is, external task information regarding an organizations’ process improvements. Often such conceptualisations are regarded to yield Tayloristic designs, which underestimate the social context. However, they possess competence in functional and technical systems design. Their strength would be increased social competence to fulfil the demand for mutual understanding, which is regarded of utmost importance in ISD (cf. Heng, Traut, & Fischer, 1999).

The designers building upon the holistic form of thought emphasise clients’ satisfaction, which ensures sustainable customer relationships and regard mutual understanding during ISD as essential between users and designers. Their strength would be increased social competence to fulfil the demand for mutual understanding, which is regarded of utmost importance in ISD (e.g., Klein & Hirschheim, 1993; Lyytinen & Ngwenyama, 1992). It seems also likely that they have competence in IS planning which aims at the improvement of organisational processes and are identified as functional, such as sales and purchasing processes, and emphasise mutual understanding. Also, they understand how to maintain customership instead of just visioning economic
Different Levels of Information Systems Designers’ Forms of Thought and Potential

gains or focusing on people’s task productivity. Besides possessing technical competence, these holistic designers would be able to consolidate definitions of formal and external work tasks into human issues. A particularly significant capability would be to understand the process of organisational learning, which is essential in order to adjust the evolving requirements during the process of ISD. Moreover, they value balanced emotional behaviour, and thus intuitively grasp the possible dangers of relying on superfluous emotional behaviour.

With respect to the humanisation of IS, a holistic conception is required in ISD. It is then assumed that the human being is actualised in intertwined physical, cognitive, emotional, social and cultural qualities, and that these qualities are fundamentally different. Without the simultaneous existence of all of the qualities, it is not possible to consider a creature as a human. Yet the qualities cannot be reduced from one quality to another, but rather need to be understood as a whole (Rauhala, 1983). Considering the human being as an actor, as a user of an IS, the whole of a human being is understood as an active subject adjoining to IS. Then the IS-user relationship consists of human action involving explicit and tacit affordances that emerge dynamically in the interaction between humans and IS. In other words, the static characteristics of humans and technology take on a new form within their intertwining activity, which is shaped according to the affordances that, on the one hand, the human substance embodies, and which, on the other hand, the properties of IS support or ignore. Consequently, understanding humans and their behaviour as users of IS requires insight into these emerging human experiences appearing within the affordances and constraints of contemporary IS and their development. Especially at present when the IS are no longer merely tools for personal and professional instrumental productivity, but also (re)constituting and mediating different social structures and practices (e.g., Orlikowski, 1992; Orlowski, 2000), IS acts as social spaces that are important growing social and cultural reference points for users and, thus also for IS designers. Design that takes into account the consequences of the form and functions of IS to users’ social qualities, such as self-identity, is indeed and necessity if contemporary IS development aims at high-quality and usable systems (Greenhill & Isomäki, 2005).

In summary, the resulting forms of thought indicate three different but associated levels of intellectual competence in conceptualising humans as users of IS. The separatist form of thought provides designers predominantly with technical perspectives and a capability for objectifying things. However, it is worth noticing that the validity of objectifying design issues is dependent on the focus of such definitions. From a human-centred perspective, valid definitions would require being theoretically sensitive to human activity and deriving abstracted conceptions from that activity rather than creating objectivist conceptualisations, which overlook humans and their behaviour. The functional form of thought focuses on external task information and task productivity, nevertheless, with the help of positive emotions. The holistic form of thought provides designers with competence of human-centred ISD, even though all the aspects of the richness of the human condition are not revealed. It seems the designers are intellectually more oriented toward designing IS for objectified, streamlined organisational processes consisting of external work tasks, and that this orientation challenges the human-centred orientations.

REFERENCES


Hedberg, B., & Mumford, E. (1975). The design of computer systems: Man’s vision of man as an integral part of the system design process. In E. Mumford, & H. Sackman (Eds.), *Human choice and computers* (pp. 31-59). *Amsterdam: North Holland.*

Different Levels of Information Systems Designers’ Forms of Thought and Potential

counting, Management and Information Technology, 9(3), 193-222.


Orlikowski, W.J. (1992). The duality of technology: Rethinking the concept of technology


kimuksen laadullisia lähestymistapoja. Oulun yliopiston kasvatustieteiden tiedekunnan opetusmonisteita ja selosteita 39 (pp. 80–107).


This work was previously published in International Journal of Technology and Human Interaction 3(1), edited by B. Stahl, pp. 30-48, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 2.15

The Think Aloud Method and User Interface Design

M.W.M. Jaspers

University of Amsterdam, The Netherlands

INTRODUCTION

Daily use of computer systems often has been hampered by poorly designed user interfaces. Since the functionality of a computer system is made available through its user interface, its design has a huge influence on the usability of these systems (Carroll, 2002; Preece, 2002). From the user’s perspective, the user interface is the only visible and, hence, most important part of the computer system; thus, it receives high priority in designing computer systems.

A plea for human-oriented design in which the potentials of computer systems are tuned to the intended user in the context of their utilization has been made (Rossen & Carroll, 2002).

An analysis of the strategies that humans use in performing tasks that are to be computer-supported is a key issue in human-oriented design of user interfaces. Good interface design thus requires a deep understanding of how humans perform a task that finally will be computer-supported. These insights then may be used to design a user interface that directly refers to their information processing activities. A variety of methodologies and techniques can be applied to analyze end users’ information processing activities in the context of a specific task environment among user-centered design methodologies. More specifically, cognitive engineering techniques are promoted to improve computer systems’ usability (Gerhardt-Powels, 1996; Stary & Peschl, 1998).

Cognitive engineering as a field aims at understanding the fundamental principles behind human activities that are relevant in the context of designing a system that supports these activities (Stary & Peschl, 1998). The ultimate goal is to develop end versions of computer systems that support users of these systems to the maximum in performing tasks in such a way that the intended tasks can be accomplished with minimal cognitive effort. Empirical research has indeed shown that cognitively engineered interfaces are considered superior by users in terms of supporting task performance, workload, and satisfaction, compared to non-cognitively engineered interfaces (Gerhardt-Powels, 1996). Methods such as the think aloud method, verbal protocol analysis, or cognitive task analysis are used to analyze in detail the way in which humans perform tasks,
mostly in interaction with a prototype computer system.

**BACKGROUND**

In this section, we describe how the think aloud method can be used to analyze a user’s task behavior in daily life situations or in interaction with a computer system and how these insights may be used to improve the design of computer systems. Thereafter, we will go into the pros and cons of the think aloud method.

**The Think Aloud Method**

Thinking aloud is a method that requires subjects to talk aloud while solving a problem or performing a task (Ericsson & Simon, 1993). This method traditionally had applications in psychological and educational research on cognitive processes. It is based on the idea that one can observe human thought processes that take place in consciousness. Thinking aloud, therefore, may be used to know more about these cognitive processes and to build computer systems on the basis of these insights. Overall, the method consists of (1) collecting think aloud reports in a systematic way and (2) analyzing these reports to gain a deeper understanding of the cognitive processes that take place in tackling a problem. These reports are collected by instructing subjects to solve a problem while thinking aloud; that is, stating directly what they think. The data so gathered are very direct; there is no delay. These verbal utterances are transcribed, resulting in verbal protocols, which require substantial analysis and interpretation to gain deep insight into the way subjects perform tasks (Deffner, 1990).

**The Use of the Think Aloud Method in Computer System Design**

In designing computer systems, the think aloud method can be used in two ways: (1) to analyze users’ task behaviors in (simulated) working practices, after which a computer system is actually built that will support the user in executing similar tasks in future; or (2) to reveal usability problems that a user encounters in interaction with a (prototype) computer system that already supports the user in performing certain tasks.

In both situations, the identification and selection of a representative sample of (potential) end users is crucial. The subject sample should consist of persons who are representative of those end users who will actually use the system in the future. This requires a clearly defined user profile, which describes the range of relevant skills of system users. Computer expertise, roles of subjects in the workplace, and a person’s expertise in the domain of work that the computer system will support are useful dimensions in this respect (Kushnirek & Patel, 2004). A questionnaire may be given either before or after the session to obtain this information. As the think aloud method provides a rich source of data, a small sample of subjects (eight to 10) suffices to gain a thorough understanding of task behavior (Ericsson & Simon, 1993) or to identify the main usability problems with a computer system (Boren & Ramey, 2000). A representative sample of the tasks to be used in the think aloud study is likewise essential. Tasks should be selected that end users are expected to perform while using the (future) computer system. This requirement asks for a careful design of tasks to be used in the study to assure that tasks are realistic and representative of daily life situations. It is recommended that task cases be developed from real-life task examples (Kushnirek & Patel, 2004).

Instructions to the subjects about the task at hand should be given routinely. The instruction on thinking aloud is straightforward. The essence
The Think Aloud Method and User Interface Design

is that the subject performs the task at hand, possibly supported by a computer, and says out loud what comes to mind.

A typical instruction would be, “I will give you a task. Please keep talking out loud while performing the task.” Although most people do not have much difficulty rendering their thoughts, they should be given an opportunity to practice talking aloud while performing an example task. Example tasks should not be too different from the target task. As soon as the subject is working on the task, the role of the instructor is a restrained one. Interference should occur only when the subject stops talking. Then, the instructor should prompt the subject by the following instruction: “Keep on talking” (Ericsson & Simon, 1993).

Full audiotaping and/or videorecording of the subject’s concurrent utterances during task performance and, if relevant, videorecording of the computer screens are required to capture all the verbal data and user/computer interactions in detail. After the session has been recorded, it has to be transcribed. Typing out complete verbal protocols is inevitable to be able to analyze the data in detail (Dix et al., 1998). Videorecordings may be viewed informally, or they may be analyzed formally to understand fully the way the subject performed the task or to detect the type and number of user-computer interaction problems.

The use of computer-supported tools that are able to link the verbal transcriptions to the corresponding video sequences may be considered to facilitate the analysis of the video data (Preece, 2002).

Prior to analyzing the audio and/or video data, it is usually necessary to develop a coding scheme to identify step-by-step how the subject tackled the task and/or to identify specific user/computer interaction problems in detail. Coding schemes may be developed bottom-up or top-down. In a bottom-up procedure, one would use part of the protocols to generate codes by taking every new occurrence of a cognitive subprocess code. For example, one could assign the code guessing to the following verbal statements: “Could it be X?” or “Let’s try X.” The remaining protocols then would be analyzed by using this coding scheme. An excerpt from a coded verbal protocol is given in Figure 1. Note that the verbal protocol is marked up with annotations from the coding scheme. Otherwise, categories in the coding scheme may be developed top-down, for example, from examination of categories of interactions from the human/computer interaction literature (Kushnirek & Patel, 2004). Before it is applied, a coding scheme must be evaluated on its inter-coder reliability.

To prevent experimenter bias, it is best to leave the actual coding of the protocols to a minimum of two independent coders. Correspondence among codes assigned by different coders to the same

---

**Figure 1. Excerpt from a coded verbal protocol for analyzing humans’ task behavior**

<table>
<thead>
<tr>
<th>Code</th>
<th>Verbal protocol segment</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPSCR04</td>
<td>How can I exit this screen?</td>
<td>Navigation problem screen04</td>
</tr>
<tr>
<td>MBT012</td>
<td>What does this button mean?</td>
<td>Meaning of button012</td>
</tr>
<tr>
<td>RTACT002</td>
<td>It has been busy a very long time</td>
<td>Response time after action002</td>
</tr>
<tr>
<td>VSSACT006</td>
<td>What is it doing now?</td>
<td>Visibility of system status after action006</td>
</tr>
<tr>
<td>MSSACT009</td>
<td>What does ‘fatal error098’ mean?</td>
<td>Meaning of system feedback after action009</td>
</tr>
</tbody>
</table>
The think aloud method and user interface design

Verbal statements must be found, for which the Kappa mostly is used (Altman, 1991).

The coded protocols and/or videos can be compiled and summarized in various ways, depending on the goal of the study. If the goal is to gain a deep insight into the way humans perform a certain task in order to use these insights for developing a computer system to support task performance, then the protocol and video analyses can be used as input for a cognitive task model. Based on this model, a first version of a computer system then may be designed. If the aim is to evaluate the usability of a (prototype) computer system, the results may summarize any type and number of usability problems revealed. If the computer system under study is still under development, these insights then may be used to better the system.

PROS AND CONS OF THE THINK ALOUD METHOD

The think aloud method, preferably used in combination with audio- and/or videorecording, is one of the most useful methods to gain a deep understanding of the way humans perform tasks and of the specific user problems that occur in interaction with a computer system. As opposed to other inquiry techniques, the think aloud method requires little expertise, while it provides detailed insights regarding human task behavior and/or user problems with a computer system (Preece, 2002). On the other hand, the information provided by the subjects is subjective and may be selective. Therefore, a careful selection of the subjects who will participate and the tasks that will be used in the study is crucial. In addition, the usefulness of the think aloud method is highly dependent on the effectiveness of the recording method. For instance, with audiotaping only, it may be difficult to record information that is relevant to identify step-by-step what the subjects were doing while performing a task, whether computer-supported or not (Preece, 2002).

Another factor distinguishing the think aloud method from other inquiry techniques is the promptness of the response it provides. The think aloud method records the subject’s task behavior at the time of performing the task. Other inquiry techniques, such as interviews and questionnaires, rely on the subject’s recollection of events afterwards. Subjects may not be aware of what they actually are doing while performing a task or interacting with a computer, which limits the usefulness of evaluation measures that rely on retrospective self-reports (Boren & Ramey, 2000; Preece, 2002). The advantage of thinking aloud, whether audio- or videotaped, as a data eliciting method includes the fact that the resulting reports provide a detailed account of the whole process of a subject executing a task.

Although using the think aloud method is rather straightforward and requires little expertise, analyzing the verbal protocols can be very time-consuming and requires that studies are well planned in order to avoid wasting time (Dix et al., 1998).

The think aloud method has been criticized, particularly with respect to the validity and completeness of the reports it generates (Boren & Ramey, 2000; Goguen & Linde, 1993).

An argument made against the use of the think aloud method as a tool for system design is that humans do not have access to their own mental processes and, therefore, cannot be asked to report on these. With this notion, verbalizing thoughts is viewed as a cognitive process on its own. Since humans are poor at dividing attention between two different tasks (i.e., performing the task under consideration and verbalizing their thoughts), it is argued that thinking aloud may lead to incomplete reports (Nisbett & Wilson, 1997).

However, this critique seems to bear on some types of tasks that subjects are asked to perform in certain think aloud studies. As Ericsson and Simon (1993) point out, in general, talking out
The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Aloud Method and User Interface Design

The Think Al oud does not interfere with task performance and, therefore, does not lead to much disturbance of the thought processes. If reasoning takes place in verbal form, then verbalizing thoughts is easy and uses no extra human memory capacity. However, if the information is nonverbal and complicated, verbalization will not only cost time but also extra human memory capacity. Verbalization of thoughts then becomes a cognitive process by itself. This will cause the report of the original task processing to be incomplete, and sometimes, it even may disrupt this process (Ericsson & Simon, 1993). Therefore, the think aloud method only may be used on a restricted set of tasks. Tasks for which the information can be reproduced verbally and for which no information is asked that is not directly used by the subject in performing the task under attention are suitable for introspection by the think aloud method (Boren & Ramey, 2000).

The fact that the experimenter may interrupt the subject during task behavior is considered another source of error, leading to distorted reports (Goguen & Linde, 1993). It has been shown, however, that as long as the experimenter minimizes interventions in the process of verbalizing and merely reminds the subject to keep talking when a subject stops verbalizing his or her thoughts, the ongoing cognitive processes are no more disturbed than by other inspection techniques (Ericsson & Simon, 1993).

The think aloud method, if applied under prescribed conditions and preferably in combination with audio- and/or videorecording, is a valuable information source of human task/behavior and, as such, a useful technique in designing and evaluating computer systems.

**FUTURE TRENDS**

The think aloud method is propagated and far more often used as a method for system usability testing than as a user requirements eliciting method. In evaluating (prototype) computer systems, thinking aloud is used to gain insight into end users’ usability problems in interaction with a system to better the design of these systems. The use of think aloud and video analyses, however, may be helpful not merely in evaluating the usability of (prototype) computer systems but also in analyzing in detail how end users tackle tasks in daily life that in the end will be computer supported. The outcomes of these kinds of analyses may be used to develop a first version of a computer system that directly and fully supports users in performing these kinds of tasks. Such an approach may reduce the time spent in iterative design of the system, as the manner in which potential end users process tasks is taken into account in building the system.

Although a deep understanding of users’ task behaviors in daily settings is indispensable in designing intuitive systems, we should keep in mind that the implementation of computer applications in real-life settings may change and may have unforeseen consequences for work practices. So, besides involving potential user groups in an early phase of system design and in usability testing, it is crucial to gain insight into how these systems may change these work practices to evaluate whether and how these systems are being used. This adds to our understanding of why systems may or may not be adopted into routine practice.

Today, a plea for qualitative studies for studying a variety of human and contextual factors that likewise may influence system appraisal is made in literature (Aarts et al., 2004; Ammenwerth et al., 2003; Berg et al., 1998; Orlikowski, 2000; Patton, 2002). In this context, sociotechnical system design approaches are promoted (Aarts et al., 2004; Berg et al., 1998; Orlikowski, 2000). Sociotechnical system design approaches are concerned not only with human/computer interaction aspects of system design but also take psychological, social, technical, and organizational aspects of system design into consideration. These approaches
take an even broader view of system design and implementation than cognitive engineering approaches—the organization is viewed as a system with people and technology as components within this system. With sociotechnical system design approaches, it can be determined which changes are necessary and beneficial to the system as a whole, and these insights then may be used to decide on the actions to effect these changes (Aarts et al., 2004; Berg et al., 1998). This process of change never stops; even when the implementation of a computer system is formally finished, users will ask for system improvements to fit their particular requirements or interests (Orlikowski, 2000).

**CONCLUSION**

The use of the think aloud method may aid in designing intuitive computer interfaces, because using thinking aloud provides us with a more thorough understanding of work practices than do conventional techniques such as interviews and questionnaires.

Until now, thinking aloud was used mostly to evaluate prototype computer systems. Thinking aloud, however, likewise may be used in an earlier phase of system design, even before a first version of the system is available. It then can be used to elicit every step taken by potential end users to process a task in daily work settings. These insights then may be used as input to the design of a computer system’s first version.

Development, fine-tuning, testing, and final implementation of computer systems take a lot of time and resources. User involvement in the whole life cycle of information systems is crucial, because only when we really try to understand end users’ needs and the way they work, think, and communicate with each other in daily practice can we hope to improve computer systems.

**REFERENCES**


The Think Aloud Method and User Interface Design


KEY TERMS

Cognitive Engineering: A field aiming at understanding the fundamental principles behind human activities that are relevant in context of designing a system that supports these activities.

Cognitive Task Analysis: The study of the way people perform tasks cognitively.

Cognitive Task Model: A model representing the cognitive behavior of people performing a certain task.

Sociotechnical System Design Approach: System design approach that focuses on a sociological understanding of the complex practices in which a computer system is to function.

Think Aloud Method: A method that requires subjects to talk aloud while solving a problem or performing a task.

User Profile: A description of the range of relevant skills of potential end users of a system.

Verbal Protocol: Transcription of the verbal utterances of a test person performing a certain task.

Verbal Protocol Analysis: Systematic analysis of the transcribed verbal utterances to develop a model of the subject’s task behavior that then may be used as input to system design specifications.

Video Analysis: Analysis of videorecordings of the user/computer interactions with the aim to detect usability problems of the computer system.

This work was previously published in Encyclopedia of Human Computer Interaction, edited by C. Ghaoui, pp. 597-602, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
INTRODUCTION

In recent years, a number of prototypical demonstrators have shown that augmented reality has the potential to improve manual work processes as much as desktop computers and office tools have improved administrative work (Azuma et al., 2001; Ong & Nee, 2004). Yet, it seems that the “classical concept” of augmented reality is not enough (see also http://www.ismar05.org/IAR). Stakeholders in industry and medicine are reluctant to adopt it wholeheartedly due to current limitations of head-mounted display technology and due to the overall dangers involved in overwhelming a user’s view of the real world with virtual information. It is more likely that moderate amounts of augmented reality will be integrated into a more general interaction environment with many displays and devices, involving tangible, immersive, wearable, and hybrid concepts of ubiquitous and wearable computing. We call this emerging paradigm ubiquitous augmented reality (UAR) (MacWilliams, 2005; Sandor, 2005; Sandor & Klinker, 2005).

It is not yet clear which UAR-based human-computer interaction techniques will be most suitable for users to simultaneously work within an environment that combines real and virtual elements. Their success is influenced by a large number of design parameters. The overall design space is vast and difficult to understand.

In Munich, we have worked on a number of applications for manufacturing, medicine, architecture, exterior construction, sports, and enter-
Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces

tainment (a complete list of projects can be found at http://ar.in.tum.de/Chair/ProjectsOverview). Although many of these projects were designed in the short-term context of one-semester student courses or theses, they provided insight into different aspects of design options, illustrating trade-offs for a number of design parameters. In this chapter, we propose a systematic approach toward identifying, exploring, and selecting design parameters at the example of three of our projects, PAARTI (Echtler et al., 2003), FataMorgana (Klinker et al., 2002), and a monitoring tool (Kulas, Sandor, & Klinker, 2004).

Using a systematic approach of enumerating and exploring a defined space of design options is useful, yet not always feasible. In many cases, the dimensionality of the design space is not known a-priori but rather has to be determined as part of the design process. To cover the variety of aspects involved in finding an acceptable solution for a given application scenario, experts with diverse backgrounds (computer science, sensing and display technologies, human factors, psychology, and the application domain) have to collaborate. Due to the highly immersive nature of UAR-based user interfaces, it is difficult for these experts to evaluate the impact of various design options without trying them. Authoring tools and an interactively configurable framework are needed to help experts quickly set up approximate demonstrators of novel concepts, similar to “back-of-the-envelope” calculations and sketches. We have explored how to provide such first-step support to teams of user interface designers (Sandor, 2005). In this chapter, we report on lessons learned on generating authoring tools and a framework for exploring interaction concepts. Finally, a report on lessons learned from implementing such tools and from discussing them within expert teams of user interface designers is intended to provide an indication of progress made thus far and next steps to be taken.

BACKGROUND

In this section, we provide an overview of the current use of UAR-related interaction techniques and general approaches toward systematizing the exploration of design options.

User Interface Techniques for Ubiquitous Augmented Reality

User interfaces in UAR are inspired by related fields, such as virtual reality (VR) (Bowman, Kruijff, LaViola, & Poupyrev, 2004), attentive user interfaces (AUIs) (Vertegaal, 2003), and tangible user interfaces (TUIs) (Ishii & Ullmer, 1997). Several interaction techniques for VR have been adapted to UAR: for example the World-in-Miniature (Bell, Höllerer, & Feiner, 2002), pinch gloves for system control (Piekarski, 2002), and a flexible pointer to grasp virtual objects that are beyond arm’s reach (Olwal & Feiner, 2003). The core idea of TUIs is to use everyday items as input and output simultaneously. This idea has also been applied to UAR: for example the World-in-Miniature (Bell, Höllerer, & Feiner, 2002), pinch gloves for system control (Piekarski, 2002), and a flexible pointer to grasp virtual objects that are beyond arm’s reach (Olwal & Feiner, 2003).
Implementing New Interaction Techniques

To develop new interaction techniques and visualizations for UAR, several software infrastructures have been created to simplify the development of new interaction techniques by programmers: distributed frameworks, dataflow architectures, user interface management systems, scene graph-based frameworks, a variety of class libraries, and finally scripting languages. A detailed discussion can be found in Sandor (2005).

For novice users, several desktop tools for authoring augmented reality content have been developed: PowerSpace (Haringer & Regenbrecht, 2002), DART (MacIntyre, Gandy, Dow, & Bolter, 2004), and MARS (Güven & Feiner, 2003). Several systems exist that follow an immersive authoring approach (Lee, Nelles, Billinghurst, & Kim, 2004; Poupyrev et al., 2001). Piekarski describes a mobile augmented reality system that can be used to capture the geometries of real objects (Piekarski, 2002). Several hybrid authoring approaches combine immersive authoring with desktop authoring (Olwal & Feiner, 2004; Zauner, Haller, Brandl, & Hartmann, 2003).

DESIGN OPTIMIZATION FOR HIGH-DIMENSIONAL DESIGN SPACES

One of the most difficult issues in designing novel interaction techniques for UAR is the wealth of criteria that are potentially involved in finding an optimal solution. We divide such criteria into three classes: criteria pertaining to the task(s) that need to be executed, the knowledge and skills of the user, and the current state-of-the-art of technology. Figure 1 illustrates the classes and their relationships.

Classes of Design Criteria

- **Task-specific criteria** are related to the requirements of specified tasks in an application. According to principles of software engineering, they are determined from scenarios and use cases, taking the environmental setting and the required technical quality into account. Yet, they may change over time due to changing work processes, which may indirectly depend on evolving technology.

- **System-specific criteria** are defined by the state-of-the-art of engineering-related parameters of sensing and display devices and computer systems. Due to evolving technology, these criteria have to be continuously re-evaluated, resulting in ever-changing optimal system configurations (Klinker et al., 1999).

- **User-specific criteria** depend on ergonomic issues and the cultural background of users—human factors and anthropology. They describe current working conditions, habits (working culture), and educational background, as well as specific user-related restrictions.
Criteria Reduction through Inter-Class Constraints

Finding an overall optimal system that works perfectly with respect to all criteria seems to be impossible. We have thus adopted the approach of selecting specific criteria of one or two classes to impose constraints on design options in other classes. In this section, we analyze the relationship between the classes of criteria from user-, system-, and task-centric specifications. Later we illustrate the exploitation of such constraints in specific examples.

The relationships between task and system requirements are described by the edge linking the task and system nodes in Figure 1. From the task-perspective, they are described as the functional and non-functional requirements of software systems. From the system-perspective, they need to be matched with the currently available technical options. Trade-offs must be made to obtain pragmatically implementable solutions—with an eye toward upcoming requirements and technical developments. Later, we will present an example of casting such trade-offs.

The relationships between task and user requirements are described by the edge linking the task and user nodes in Figure 1. This case does not involve any considerations of currently available technology. Thus, options that are discussed here should hold true now, as well as 100 years ago or 100 years in the future. They are analyzed by disciplines such as task analysis and system ergonomics (Bubb, 1993). Yet, they can provide significant constraints upon today’s technically achievable system configurations. We will present an example of analyzing how a user (car designer) physically behaves with respect to a number of tasks geared toward analyzing and comparing different automotive designs.

The relationships between system and user are described by the edge linking the system and user nodes in Figure 1. From the user-perspective, they are described as usability criteria, evaluating how users perform, given a specific technical system—in comparison to other technical options. From the system-perspective, they describe user requirements that need to be satisfied with currently available technical means. Later, we show an example of how a specific technical device can be evaluated with respect to specific physical user skills in using such a device.

Dealing with Ill-Defined Design Spaces

By applying inter-class constraints on a user interface, the design space can often be reduced considerably. The next step in our proposed design process is to explore the reduced design space with interactive tools that encourage collaboration. In this section, we first give the rationale for our interactive tools. Then, we proceed by highlighting the problems that occur when using this approach. Finally, we give an outlook of how we elaborate on these concepts within this chapter.

To further explore the design space, collaboration between researchers with different backgrounds is imperative to yield a solution that is well balanced according to our three main classes: user, task, and system. Thinking about this problem led to the invention of a new development process: Jam Sessions. The name Jam Sessions was inspired by the spontaneous collaboration of Jazz musicians that is also named Jam Sessions. However, in our case we collaborate on user interface elements, instead of music. In Jam Sessions, development takes place at system runtime, next to a running system. This allows playful exploration of user interface ideas. Our experience with Jam Sessions was first presented in (MacWilliams et al., 2003); we have already discussed these from a software engineering (MacWilliams, 2005) and user interface (Sandor, 2005) perspective.

To support this development process, interactive tools for novices are an important ingredient, since they foster the interdisciplinary collabora-
Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces

...tion with other researchers. Desirable would be a set of generic tools that can be applied in all Jam Sessions—independent of the user interface to be developed. Although we have achieved this for programmers, for novices this is yet an unsolved problem. We go in line with several other research tools that allow modifying only a quite limited amount of user interface functionality. Since these tools are customized toward the user interface that has to be built, most projects require writing new tools. Thus, a sophisticated software infrastructure that quickly allows building new tools is very useful.

The final section of this chapter describes a complex user interface that we have designed in Jam Sessions. Additionally, we first describe briefly our software infrastructure and elaborate on the tools that we have created for this project.

PROJECTS USING INTER-CLASS CONSTRAINTS

This section presents three examples of analyzing design options by exploring inter-class constraints.

PAARTI

In the PAARTI project (practical applications of augmented reality in technical integration), we have developed an intelligent welding gun with BMW that is now being used on a regular basis to weld studs in the prototype production of cars (Echtler et al., 2003). It exemplifies the systematic exploitation of constraints between task and system criteria.

The task was to assist welders in positioning the tip of a welding gun with very high precision at some hundred predefined welding locations on a car body. The main system design issue was to find an immersive solution with maximal precision. An AR-based system would need a display (D), a tracking sensor (S), and some markers (M) that needed to be installed in the environment on the user or on the welding gun in a manner that would yield maximal precision. As a fourth option, we considered the case that one of the objects (esp.: markers) would not be necessary at all. The result was the definition of a 3-dimensional design space, $S \times M \times D$, with each dimension spanning a range of four options. In total, there were $4^3 = 64$ solutions that needed to be considered.

According to an analysis of all options, the highest precision could be achieved by using an outside-in tracking arrangement with sensors placed in the welding environment and markers attached to the welding gun. A small display was attached to the welding gun. The visualization used a notch and bead metaphor of real guns, consisting of several concentric rings. A sphere was positioned three-dimensionally at the next welding location. Welders were requested to capture the sphere within the concentric rings by moving the gun (and the display) to the appropriate location (see Figure 2(a)).

The left diagram of Figure 2(b) shows the optimal setup. The center diagram of Figure 2(b) indicates the design space of $S \times M = 16$ options of placing markers and sensors in the environment, with the fixed display dimension D of placing the display on the welding gun. The right diagram presents the analysis pertaining to the achievable precision for this setup. This setup clearly outperforms the classical AR-arrangement involving a head-mounted display (Echtler et al., 2003).

FataMorgana

In the FataMorgana project, we have developed an AR-based prototypical demonstrator for designers at BMW, helping them compare real mock-ups of new car designs with virtual models (Klinker et al., 2002). The rationale for building this system was that, although the importance of digital car models is increasing, designers have not yet committed wholeheartedly to a VR-based approach but rather prefer relying on physical
mock-ups. One of the reasons may be that special viewing arrangements such as projection walls do not permit people to view the digital models within a real environment. AR can help alleviate this problem by placing virtual cars next to real (mock-up) cars.

Here we present this project as an example of a systematic analysis of the relationships between tasks and user actions. The underlying thesis is that users (designers) behave in specific ways in order to achieve tasks. If a system is expected to support users in achieving their tasks, it has to be designed to function well within the range of typical actions performed by the user. To this end, we have subdivided the task in a set of different approaches and asked a designer to act out each of these tasks within the real car presentation environment. We recorded the designer’s motions with a camera that was attached to his head (Figure 3).
Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces

Figure 3. Five different tasks resulting in different user head motions (© 2002 IEEE, Klinker et al., 2002)

- **Turning**: The car designer remains in a fixed location and looks at the car rotating on a turn table.
- **Overview**: The car designer performs an overview evaluation of the car, by walking around and turning his head to change the lighting conditions.
- **Detail**: The car designer focuses on a specific detail of the car, such as a character line on the side of the car or the shape of the front spoiler.
- **Discuss**: The car designer discusses the car under evaluation with a colleague.
- **Compare**: The car designer compares two cars, for example, an existing car and a new design.

For each scenario, we determined the typical angular range of head rotations, as well as the range of positional changes. Combined with a projection of the field of view onto the environment, this gave us an indication how markers had to be laid out in the room in order to guarantee that enough of them were clearly visible during all user actions.

**Monitoring Tool for Determining Usage Patterns of Novel Interaction Techniques**

We have developed a monitoring tool (Kulas et al., 2004) to evaluate the usability of novel input techniques and devices. The monitoring tool allows us to systematically analyze relationships between user and system criteria, analyzing whether a system is well tailored to the physiological and cognitive skills of its users.

We have used the monitoring tool to evaluate the usability of a novel input device called TouchGlove that was developed at Columbia University (Blasko & Feiner, 2002). It consists of a touch-sensitive plate (similar to a touchpad in a laptop) that is attached to the center of a user’s palm. It is sensitive to single-finger input, measuring 2D location and pressure. In the evaluation setup, we have compared two techniques of using
Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces

the TouchGlove to select items from a menu. In the first case, users were asked to make a linear gesture with their fingertip on the TouchGlove to select items from a regular pull-down menu. In the second case, the TouchGlove was combined with a gyroscope to select items from a pie menu. Users were asked to rotate their hands around their wrists, generating only a tapping signal on the touchpad to signal start and end of the gesture rotating the pie menu.

During a usability evaluation, the user is placed at a suitable distance from a usability engineer. The engineer enters observations into a usability logging system and also monitors what the user actually sees on screen. Simultaneously, he also monitors real-time visualizations of measured usability data. The tool provides immediate feedback during an interactive tryout session, thereby supporting Jam Sessions.

INTERACTIVE TOOLS FOR COLLABORATIVE DESIGN SPACE EXPLORATIONS

This section presents our tools for supporting Jam Sessions. First, we give a brief overview of our tools. Second, we present an interdisciplinary research project, CAR, which uses them. We close with a description of the underlying real-time development environment.

Overview of Tools

To support Jam Sessions, we have created a toolbox of lightweight and flexible tools. They form the basic building blocks which user interface development teams can use to generate, experience, and test their novel interaction techniques.

The tools use AR, TUI, and WIMP interaction paradigms and are designed to support a number of tasks. The first task focuses on monitoring the user. The second task involves the configuration of dataflow networks. UAR systems need to communicate in real-time with many sensing and display devices, requiring a distributed system approach. A dataflow network connects such devices and components. We provide tools that allow modifying these dataflow graphs during runtime. Another task is related to the adjustment of dialog control, i.e., the control of the high-level behaviour of a user interface. Tools that enable developers to specify dialog control quickly speed up the development process significantly. The final task involves the creation of context-aware animations. Conventional animations have time as the only parameter that changes the appearance of graphical elements. However, for mobile systems a variety of research projects (e.g., a context-aware World-in-Miniature [Bell et al., 2002]) have explored animations that change their appearance according to context.

We have developed six tools, T1–T6, in support of these tasks. T1 collects and evaluates usability data during system runtime. T2 uses an augmented reality visualization to shows a user’s visual focus of attention in a combination of head and eye tracking (Novak et al., 2004) (see Figure 9). T3 is a graphical editor, DIVE to adjust dataflow networks (MacWilliams et al., 2003; Pustka, 2003) (see Figure 10[a]). T4 is an immersive visual programming environment (Sandor, Olwal, Bell, & Feiner, 2005). T5 is a User Interface Controller Editor, UIC, to graphically specify dialog control by composing Petri nets (Hilliges, Sandor, & Klinker, 2004) (see Figure 10[b)). T6 is a collection of tools to experiment with context-aware mobile augmented reality user interfaces.

Figure 4 classifies our tools with respect to the user interface paradigms they employ and the tasks they address (Sandor, 2005). It shows that we sometimes developed several tools, addressing the same task, using different interaction paradigms. This reflects our goal of exploring and comparing design options for our own tools as much as for the interaction techniques that will be developed with them.
Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces

CAR

CAR is an industry-sponsored multi-disciplinary project to investigate issues pertaining to the design of augmented reality user interfaces in cars. CAR has used most of the tools T1-T6 to investigate several user interface questions.

Motivation

In CAR, we have investigated a variety of questions: How can information be presented efficiently across several displays that can be found in a modern car (e.g., the dashboard, the board computer, and heads-up displays (HUDs))? How can we prevent that information displayed in a HUD is blocking the driver’s view in crucial situations? Since a wealth of input modalities can be used by a car driver (tactile, speech, head and hand gestures, eye motion), which modalities should be used for which tasks?

In a multi-disciplinary UI design team, we have discussed, for example, how to present a navigation map on a HUD. Where should it be placed? How large should it be? What level of detail should it provide? Should it be a two-dimensional map or a tilted view onto a three-dimensional environmental model (WIM)? If so, which viewing angle should be selected? Will the angle, as well as the position of the WIM and the size and zoom factor adapt to sensor parameters, such as the current position of the car while approaching a critical traffic area in a town?
Physical Setup

We have set up a simulator for studying car navigation metaphors in traffic scenes (Figure 5). It consists of two separate areas: a simulation control area (large table with a tracked toy car) and a simulation experience area (person sitting at the small table with a movable computer monitor in the front and a stationary large projection screen in the back). In the simulation control area, members of the design team can move one or more toy cars on the city map to simulate traffic situations, thereby controlling a traffic simulator via a tangible object. The simulation experience area represents the cockpit of a car and the driver. The picture projected on the large screen in the front displays the view a driver would have when sitting in the toy car. The monitor in front of the driver provides a mock-up for the visualizations to be displayed in a HUD. Further monitors can be added at run-time, if more than one view is needed.

The room is equipped with an outside-in optical tracking system (http://www.ar-tracking.de). The cameras track the toy car, the computer monitor, and the user (simulating a car driver). Each tracked object is equipped with a marker consisting of a rigid, three-dimensional arrangement of reflective spheres.

Information is presented on several devices and surfaces in the room: A projector at the ceiling projects a bird’s eye view of a city onto the large, stationary table on the right. Another projector presents the current, egocentric view of a virtual car driver sitting in the toy car on the large screen at the front wall. A third, location-dependent visualization of the driving scenario is shown on the mobile computer monitor—our substitute for a HUD.

The system provides tools for a team of design experts with diverse backgrounds to jointly explore various options to present a map (Figure 6).
Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces

Controlling the Context-Aware Adjustment of a Navigation Map

It is not yet clear how navigational aids are best presented within a driver’s field of view. In the CAR project, we have experimented with various options of placing and orienting a map in a HUD.

Figure 7 shows how our system provides designers with a tangible object—a plate—that is correlated with the orientation (tilt) and zoom of a 3D map on the HUD. When the user moves the tangible plane, the 3D map is turned and zoomed accordingly on the HUD. Figure 6(a) shows a member of the design team experiment with different map orientations.

The position and size of a map in a HUD may have to depend on various parameters that depend on the driving context, such as the current position of the car relative to its destination, the driver’s viewing direction, and immanent dangers in the environment. Interface designers need to explore schemes for the display system to automatically adapt to context parameters. Figure 8 shows an interactive sketching tool for designers to describe functional dependencies between context parameters and display options. Figure 9 shows first steps toward using tracked head and eye motions to provide a context-dependent interaction scheme (Novak et al., 2004). Figure 6(a) shows the head and eye tracking device. We are in the process of analyzing context-dependent information presentation further. First user studies of selected issues are presented in Tönnis, Sandor, Klinker, Lange, and Bubb (2005).

Real-Time Development Environment for Interaction Design

The tools presented in earlier sections were geared toward the immediate use by non-programming user interface experts. They mainly address the customization of a set of functionalities and filters, linking context measurements to information presentation schemes. In order to add new functionality to a system, the development team must also be able to modify the underlying network of components, and its dataflow scheme. Tools T3 and T5 provide such support.

All system configuration tools are based on DWARF (distributed wearable augmented reality framework) (Bauer et al., 2001) and AVANT-
Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces

Figure 7. Tangible interaction for adjustment of a three-dimensional map

Figure 8. Sketching the context visualization function: (a) Staircase function; (b) linear function

Figure 9. Attentive user interface, visualizing a driver’s eye and head motions: (a) The DWARF’s Interface Visualization Environment for managing distributed components; (b) The User Interface Controller for specifying dialog control

GUADE (Sandor, 2005; Sandor & Klinker, 2005). DWARF is the underlying infrastructure that connects a set of distributed components. AVANTGUADE is composed of DWARF components that address the specific requirements for user interfaces in UAR.

DWARF’s interactive visualization environment (MacWilliams et al., 2003) (tool T3, Figure 10[a]) enables developers to monitor and modify the dataflow network of distributed components. However, since this requires substantial knowl-
Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces

The core component of AVANTGUARDE is a Petri net-based dialog control management system (Hilliges et al., 2004) (Tool T5, Figure 10[b]). We have developed a visual programming environment that eases the modification of the Petri nets (and accordingly the user interface) during system runtime. However, it is still too difficult to use for non-programming design experts, since understanding Petri nets requires knowledge in computer science.

CONCLUSION

In PAARTI and FataMorgana, we have learned that the reduction of criteria through inter-class constraints is a valuable approach for designing user interfaces. The crucial issue of this method is to determine the most important constraints by talking with domain experts. The subsequent systematic design space exploration is straightforward.

We have presented an example for the inter-class constraint of user and system: the evaluation of the TouchGlove input device. In this project, we have observed the importance of immediate feedback through interactive tools. Our first prototype of the TouchGlove had a loose contact. While we conducted the usability study with the first user, we immediately spotted the problem and solved it. This saved us a lot of valuable time.

Our tool-based approach for further design space explorations has been applied successfully in several projects. The idea of providing user interface developers with a toolbox of flexible, lightweight tools seems feasible. However, one
problem has to be pointed out: when creating a variety of tools, a supporting real-time development environment is imperative. Otherwise, too much development time has to be allocated to tool creation—leaving little time for the actual use of the tools. In this respect, we have successfully built our tools on top of DWARF and AVANTGUARDE.

The combination of tools with different user interface paradigms turned out to be a valuable idea. We have made two important observations: first, there seems to be a trade-off between ease of use for a tool and the complexity of results that can be accomplished with it. WIMP tools can be used to model more complex interactions, whereas ease of use is greater with tools that have a tangible user interface or an augmented reality user interface. Second, the combination of tools with different paradigms opens new possibilities of interaction design that would not be possible with tools employing a single paradigm. Interaction designers are typically not fluent in complex programming tasks, so their involvement with easy to use tools yields important benefits.

Ideally, it would be enough to create one generic tool that novices can use to explore the design space of UAR user interfaces. Our first prototype toward this goal has been published in Sandor et al. (2005).

This tool seems to be very easy to use, as it employs only direct manipulation of real world objects—no conventional programming is required at all. However, the ceiling of the tool (i.e., what can be achieved with it) is quite low, since our system supports only a fixed, and very limited number of operations. We are exploring how we can extend it to allow users to specify new operations at runtime. While we anticipate using programming-by-demonstration to address a carefully planned universe of possibilities, supporting arbitrary operations through demonstration and generalization is an open problem.

The CAR project also showed us that for design space explorations, rapid prototyping is more important than realism for finding new interaction techniques. Though, for the thorough evaluation of these new concepts, formal usability studies within a realistic environment are still necessary. We have conducted a first study in this respect (Tönnis et al., 2005).

REFERENCES


Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces


Lessons Learned in Designing Ubiquitous Augmented Reality User Interfaces


This work was previously published in Emerging Technologies of Augmented Reality: Interfaces and Design, edited by M. Haller, B. Thomas, and M. Billinghurst, pp. 218-235, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 2.17
Social Network Structures in Open Source Software Development Teams

Yuan Long
Colorado State University-Pueblo, USA

Keng Siau
University of Nebraska-Lincoln, USA

ABSTRACT
Drawing on social network theories and previous studies, this research examines the dynamics of social network structures in open source software (OSS) teams. Three projects were selected from SourceForge.net in terms of their similarities as well as their differences. Monthly data were extracted from the bug tracking systems in order to achieve a longitudinal view of the interaction pattern of each project. Social network analysis was used to generate the indices of social structure. The finding suggests that the interaction pattern of OSS projects evolves from a single hub at the beginning to a core/periphery model as the projects move forward.

INTRODUCTION
The information system development arena has seen many revolutions and evolutions. We have witnessed the movement from structured development to object-oriented (OO) development. Modeling methods, such as data flow diagram and entity relationship diagram, are facing new OO modeling languages, such as the unified modeling language (UML) (see Siau & Cao, 2001; Siau, Erickson, & Lee, 2005; Siau & Loo, 2006) and OO methodologies, such as unified process (UP). The latest development includes agile modeling (see Erickson, Lyytinen, & Siau, 2005), extreme programming, and OSS development. While many of these changes are related to systems development paradigms, methodologies, methods, and techniques, the phenomenon of OSS development entails a different structure for software development teams.
Unlike conventional software projects, the participants of OSS projects are volunteers. They are self-selected based on their interests and capability to contribute to the projects (Raymond, 2000). In addition, the developers of OSS projects are distributed all around the world. They communicate and collaborate with each other through the Internet, using e-mails or discussion boards. Therefore, effective and efficient communication and collaboration are critical to OSS success. However, little empirical research has been conducted to study the underlying interaction pattern of OSS teams, especially the dynamics of the social network structures in OSS development teams. To fill this gap, this study examines the evolvement of social structure in OSS teams. The study contributes to the enhancement of the understanding of OSS development, and provides foundation for future studies to analyze the antecedents and consequences of social networks in the OSS context.

The remainder of the paper is structured as follows. First, prior studies on social network structures in OSS teams are reviewed. Second, theories related to social structure and social network theory are discussed. Third, the research methodology is presented, and the research results are reported. Next, discussions of the results, the limitations, and the implications are provided. The paper concludes with suggestions for future research.

LITERATURE REVIEW

The phenomenon of OSS development has attracted considerable attention from both practitioners and researchers in diverse fields, such as computer science, social psychology, organization, and management. Because of the multifaceted nature of OSS, researchers have investigated OSS phenomenon from varied perspectives. For example, focusing on technical perspective, researchers studied issues such as OSS development methodology (e.g., Jørgensen, 2001) and coding quality (e.g., Stamelos, Angelis, Oikonomu, & Bleris, 2002). Based on social psychology, researchers investigated individual motivation (e.g., Hann, Robert, & Slaughter, 2004), new developers (Von Krogh, Spaeth, & Lakhani, 2003), the social network (e.g., Madey, Freeh, & Tynan, 2002), and the social structure (e.g., Crowston & Howison, 2005). In terms of organizational and managerial perspective, researchers examined knowledge innovation (e.g., Hemetsberger, 2004; Lee & Cole, 2003, Von Hippel & Von Krogh, 2003) and the governance mechanism (e.g., Sagers, 2004).

An OSS development team is essentially a virtual organization in which participants interact and collaborate with each other through the Internet. Compared to conventional organizations, the structure of virtual organizations is decentralized, flat, and nonhierarchical (Ahuja & Carley, 1999). However, some researchers challenge the belief (e.g., Crowston & Howison, 2005; Gacek & Arief, 2004; Mockus, Fielding, & Herbsleb, 2000; Mockus, Fielding, & Herbsleb, 2002; Moon & Sproull, 2000). They argue that the social structure of OSS projects is hierarchical rather than flat, like a tree (Gacek & Arief, 2004) or an onion (Crowston & Howison, 2005). The social structure of OSS teams directly influences the collaboration and the decision-making process and further affects the overall performance of the teams as well as individuals’ perception of belonging and satisfaction. Therefore, one wonders what form of social structure might be present in the OSS development and what type of structure will emerge—centralized or decentralized, hierarchical or nonhierarchical, onion-like or tree-like, or a combination of the above depending on certain specific situations?

A social network, as stated by Krebs and Holley (2004), is generally built in four phases, each with its own distinct topology (as shown in Figure 1).
1. scattered clusters,
2. single hub-and-spoke,
3. multihub small-world network, and
4. core/periphery.

Most organizations start from isolated and distributed clusters (Krebs & Holley, 2004). Then an active leader emerges and takes responsibility for building a network that will connect the separate clusters. However, this single-hub topology is fragile. With more participants entering the group, the leader changes his/her role to a facilitator and helps to build multiple hubs, which is stage three. The core/periphery model, the last stage, is the most stable structure. In the core/periphery model, the network core encompasses key group members who are strongly connected to each other, while the periphery contains members who are usually weakly connected to each other as well as to the core members. With the general network building phases in mind, one can argue that OSS projects may follow the same four stages (i.e., scattered clusters, single hub-and-spoke, multihub small-world network, and core/periphery model). But is that true for OSS projects? How does the social structure of OSS teams evolve over time?

Our research addresses the following two questions:

1. What is the social structure of OSS teams?
2. How does the social structure evolve over time?
THEORETICAL FOUNDATION

Social Structure and Social Interaction

Social structure, as suggested by Schaefer and Lamm (1998), refers to the way in which society is organized into predictable relationships. Social structure can be considered in terms of three aspects—actors, their actions, and their interactions. The social actor is a relatively static concept addressing issues such as roles, positions, and statuses. Individual actors are embedded in the social environment and, therefore, their actions are largely influenced by the connections between each other. Social interaction is generally regarded as the way in which people respond to one another. These interaction patterns are to some extent independent of individuals. They exert a force that shapes both behavior (i.e., actions) and identity (i.e., actors) (Schaefer & Lamm, 1998).

Research on social interaction focuses on how individuals actually communicate with each other in group settings. These studies address issues such as the interaction patterns, the underlying rules guiding interaction, the reasons accounting for the way people interact, and the impacts of the interaction patterns on the individual behavior and the group performance. These issues begin by questioning what might be the interaction pattern in a specific social setting and that addresses our research question—understanding social interaction of OSS project teams.

Social Network Theory

Social network theory focuses on studying actors as well as their relationships in specific social settings. Network theory is analogous to systems theory and complexity theory. It is an interdisciplinary theory stemming from multiple traditional fields, including psychology, which addresses individuals’ perception of social structure; anthropology, which emphasizes social relationships; and mathematics, which provides algorithms (Scott, 2000).

Based on the view of social network, the world is composed of actors (also called nodes) and ties between them. The ties can represent either a specific relationship (such as friendship and kinship) between a pair of actors or define a particular action which an actor performs. Different kinds of ties specify different networks and are typically assumed to function differently. For example, the ties in a family network are distinctive from those in a working network, and the centrality in the “who loves whom” network obviously has different meaning than the centrality in the “who hates whom” network.

Social network theory is based on the intuitive notion that the social interaction patterns are essential to the individuals who reflect them. Network theorists believe that how individuals behave largely depends on how they interact with others and how they are tied to a social network. Furthermore, besides individual behavior, network theorists believe that the success or failure of societies and organizations often depends on the internal interaction pattern (Freeman, 2004).

Besides the theoretical essence, social network theory is also characterized as a distinctive methodology encompassing techniques for data collection, statistical analysis, and visual representation. This approach is usually called social network analysis and will be discussed in the research methodology section. This paper draws on the social network theory to study the interaction pattern of OSS development project.

RESEARCH METHODOLOGY

Social Network Analysis

Social network analysis is used in our study to investigate the interaction pattern of the OSS development process. Social network analysis
focuses on uncovering the interaction pattern of interdependent individuals (Freeman, 2004). Through a structural analysis of a social network diagram, a map depicting actors as well as the ties that connect them, social network analysis can reveal the patterns of relationships and the relative position of individuals in a specific social setting. This approach has been effectively used in organizational research, social support, mental health, and the diffusion of information (Freeman, 2004).

Social network analysis is used in our study for two primary reasons. First, the purpose of social network analysis fits our research objective. Social network analysis aims to analyze the relationship among a set of actors instead of their internal attributes. Our research aims to reveal the interaction pattern of OSS project teams. Therefore, social network analysis is helpful in answering our research questions.

Second, the rich interactive data extracted from OSS projects presents a “gold mine” for social network analysis. Social network analysis is grounded in the systematic analysis of empirical data. However, there is usually a lack of convenient and objective resources from which to draw the links (i.e., relationships) among actors. Most OSS projects have online mailing lists, forums, and tracking systems that are open to public, thus providing a rich set of longitudinal data. Based on these public data, researchers are able to capture input data sets for social network analysis.

**Longitudinal Data**

Because we are interested in studying how the interaction pattern of OSS projects evolves over time, cross-sectional observations of interaction networks are not sufficient. Cross-sectional observations of social networks are snapshots of interactions at a point in time and cannot provide traceable history, thus limiting the usefulness of the results. On the other hand, longitudinal observations offer more promise for understanding the social network structure and its evolvement. In this study, we extracted longitudinal data on OSS projects.

**Case Selection**

OSS projects were selected from the SourceForge\(^1\), which is the world’s largest Web site hosting OSS projects. SourceForge provides free tools and services to facilitate OSS development. At the time of the study, it hosted a total of 99,730 OSS projects and involved 1,066,589 registered users (This data was retrieved on May 4, 2005). Although a few big OSS projects have their own Web sites (such as Linux), SourceForge serves as the most popular data resource for OSS researchers.

Following the idea of theoretical sampling (Glaser & Strauss, 1967), three OSS projects were selected from SourceForge in terms of their similarities and differences. Theoretical sampling requires theoretical relevance and purposes (Orlikowski, 1993). In terms of relevance, the selection ensures that the interaction pattern of OSS projects over time is kept similar. Therefore, the projects that are selected have to satisfy two requirements. First, the projects must have considerable interaction among members during the development process. All three projects had more than 10 developers, and the number of bugs reported was more than 1,000. Second, since we are interested in the interaction over time, the projects must have a relatively long life. In our case, all three projects were at least three years old.

In addition to similarities, differences are sought among cases because the study aims to study interaction patterns of various OSS projects. Therefore, the three projects differ on several project characteristics, such as project size, project type, and intended audience. These differences enable us to make useful contrasts during data analysis.

The Table 1 summarizes the three projects with a brief description.
Data Collection and Analysis

Social network analysis can be divided into the following three stages (Borgatti, 2002).

1. Data collection. In this stage, researchers collect data, using surveys and questionnaires, or from documents and other data resources, and generate input data sets for social network analysis.

2. Statistical analysis. Based on mathematics algorithms, this stage generates network indices concerning group structure (such as centralization and density) as well as individual cohesion (such as centrality and bridges).

3. Visual representation. This stage employs network diagrams to show the interaction structure as well as the position of specific actors.

First is the data collection. The data were collected in April 2005 from SourceForge.net. Data were extracted from the bug tracking system of each project. We chose the bug tracking system as the primary data resource for three reasons. First, open source software is characterized as peer review of open codes. Raymond (1998) proposed the “Linus’ law” in his well-known essay *The Cathedral and the Bazaar*— “Given enough eyeballs, all bugs are shallow.” Therefore, the bug system can be viewed as the representative of open source spirit. Second, compared to other development activities, such as patch posting and feature request, the bug-fixing process is the most active procedure to illustrate the close

---

**Table 1. Summary of three projects**

<table>
<thead>
<tr>
<th>Description</th>
<th>Net-SNMP allows system and network managers to monitor and manage hosts and network devices.</th>
<th>Compiere is a smart ERP+CRM solution covering all major business areas—especially for small-medium enterprises.</th>
<th>JBoss is a leading open source Java application server. After Linux and Apache, it is the third major open source project to receive widespread adoption by corporate IT.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarities</td>
<td>Bug reports (more than 1,000 bugs) 1,361</td>
<td>1,695</td>
<td>2,296</td>
</tr>
<tr>
<td>Development duration (more than 3 years)</td>
<td>55 months (registered on 10/2000)</td>
<td>47 months (registered on 6/2001)</td>
<td>50 months (registered on 3/2001)</td>
</tr>
<tr>
<td>Differences</td>
<td>Software type</td>
<td>Internet, network management</td>
<td>Enterprise: ERP+CRM</td>
</tr>
<tr>
<td>Group size (number of developers)</td>
<td>Small (14)</td>
<td>Median (44)</td>
<td>Large (75)</td>
</tr>
<tr>
<td>Intended audience</td>
<td>Developers, system administrators</td>
<td>Business</td>
<td>Developers, system administrators</td>
</tr>
</tbody>
</table>
Social Network Structures in Open Source Software Development Teams

A Web spider program, which is based on the work of Crowston and Howison (2005) with necessary revision, was used to download the bug tracking Web pages from the project Web site. After that, a Web parsing program was developed to analyze the Web pages. The interaction data was extracted from the bug tracking Web pages month-by-month, starting from the date the project was registered until the date the data was downloaded for this study. The output of this stage is a social matrix describing the interaction among users. Figure 2 shows an example of such a social matrix for an OSS project. In the matrix, each row or column represents a distinctive participant, which is identified by a unique SourceForge user identity. The values of cells indicate the degree of the interaction between each pair of participants, which is counted by the amounts of messages that participant A (i.e., row A) replied to participant B (i.e., column B).

Second is the statistical analysis. Our study focuses on two important and distinctive properties of network structure—group centralization and core/periphery fitness. Ucinet, which was developed by Borgatti, Everett, and Freeman (2002), was used to calculate these two properties.

Group centralization, as suggested by Wasserman and Faust (1994), refers to the extent to which a network revolves around a single center. A typical case of centralized structure is a “star” network. Group centralization can be viewed as a rough measure of inequity between individual actors, and the variability and dispersion of the interaction pattern.

The other property is core/periphery fitness. It measures the extent to which the network is close to a perfect core/periphery structure. The core/periphery structure depicts a dense, connected group
surrounded by a sparse, unconnected periphery. The opposite structure is clique, which represents a structure of multiple subgroups, each with its own core and peripheries (Borgatti, 2002).

Finally is the visual representation. We used Ucinet (Borgatti et al., 2002) to draw the interaction networks for each of the three projects.

**RESEARCH RESULTS**

**Snapshots of the Three Projects**

Monthly data were extracted from the bug tracking system of each project. To illustrate the trend of interaction pattern, we provide three snapshots for each project (see Figures 3-5)

Table 2 summarizes the relevant network characteristics of each project. In addition to the group centralization and core/periphery fitness, we also report other network characteristics, such as density, average distance, and distance-based cohesion. Density depicts how “close” the network looks, and it is a recommended measure of group cohesion (Blau, 1977; Wasserman & Faust 1994). The value of density ranges from 0 to 1. Average distance refers to average distance between all pairs of nodes (Borgatti, 2002). Distance-based cohesion takes on values between 0 and 1—the larger the values, the greater the cohesiveness.

Looking at the statistical results and the network plots, we can observe the following.

First, the evolvement of interaction patterns of the three projects reveals a general trend. As shown in the network plots (i.e., Figures 3-5), the interaction pattern develops from a centralized one with a single (sometimes dual) hub with several distributed nodes, to a core/periphery
structure that has a core (a group of core developers) together with several hangers-on (periphery). Intense interactions exist within the core (among several core developers) and between each core member and his/her periphery. However, only loose relationships exist among peripheries. This pattern suggests a layer structure (i.e., core with its periphery) instead of a complete flat one with equal positions across all the members.
Second, although the interaction patterns of the three projects share some commonalities, their exact shapes are different. The shape of Net-SNMP (as shown in Figure 3) is more like a typical core/periphery compared to the other two. Compiere (as shown in Figure 4) keeps two cores, and the shape looks like a dumbbell. Jboss (as shown in Figure 5), which is the largest project among the three, maintains a more complex structure that shows multiple layers instead of just one core with the rest as peripheries (e.g., Net-SNMP).

Third, as time goes by, the group centralization decreases across the three projects, showing a trend that moves from a centralized structure to a decentralized structure irrespective of project sizes (the three projects with different project sizes are shown in Table 1), project types, and intended audience.

Fourth, the indices of core/periphery fitness of each project fluctuate slightly but maintain a relatively high value (larger than 0.5 on average). However, no observable trend was found across projects.

Fifth, since each project has a relatively large group (i.e., more than 100 actors including all the registered users), the values of density are relatively low with little variation. Therefore, density is not appropriate for comparing the projects.

From the snapshots, we observed the following trend. First, the OSS interaction network evolves into a core/periphery structure. Second, group centralization decreases over time. Third, core/periphery fitness stays relatively stable. To verify our observations, we used longitudinal data generated from the bug tracking systems to analyze the evolution of interaction pattern (discussed in the following section).
Social Network Structures in Open Source Software Development Teams

Table 3. Group centralization and core/periphery fitness based on longitudinal data

<table>
<thead>
<tr>
<th></th>
<th>Group centralization</th>
<th>Core/periphery fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net-SNMP</td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Compiere</td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>JBoss</td>
<td><img src="image" alt="Graph" /></td>
<td><img src="image" alt="Graph" /></td>
</tr>
</tbody>
</table>

**Group Centralization and Core/Periphery Fitness**

Table 3 shows the values of both group centralization and core/periphery fitness over time based on the monthly interaction data. For each figure, the Y-axis indicates the social structure indices (i.e., group centralization or core/periphery fitness), and the X-axis reflects the time dimension.

Two primary observations can be made based on the statistical analysis.

First, the group centralization shows a decreasing trend across the three projects. This observation indicates that as OSS projects progress, the social network structure evolves from centralized to decentralized and then stabilizes. Also, the three figures suggest no substantial differences in the trend among the three projects.

Second, the core/periphery index is maintained at a relatively stable level for each project over time. In addition, the average fitness value stays relatively high for each project (larger than 0.5), indicating a closeness to a perfect core/periphery structure.

Besides a holistic view of network structure for OSS projects, the results also reveal other interesting findings. For example, by examining the core members over time, we found a relatively stable core for each project. The cores are usually project developers and administrators. This observation further demonstrates the existence of strong and stable core as well as loose hangers-on in OSS projects.
DISCUSSION

This research uses the longitudinal data of three OSS projects selected from SourceForge to study the social network structures of OSS teams. The purpose of this study is to investigate the evolution of interaction patterns of OSS project teams. The research results suggest a decrease of group centralization over time and a tendency of core/periphery structure in OSS project teams.

The network plots (as shown in Figures 3-5) indicate a layer structure instead of a flat one as suggested by earlier literature. The interaction pattern evolves from a single hub to a core/periphery structure. As the number of participants increases, a core with only one person (who may be the starter/initiator of the project) cannot satisfy the increasing requirements of development and communication. Therefore, other developers or active users join the core to serve as key members of the project. This results in a more stable structure, and the project is less dependent on a single leader.

With the growth of a software project, more people are attracted to the project. The original leader may not be able to solve all the technical problems encountered in the development process. Each key member has his/her specialty, is responsible for solving relevant problems, and has his/her own periphery in the network plot. Although there are multiple peripheries in the project, collaboration among key members in the project is vital. This phenomenon of distribution and collaboration can be viewed as a critical success factor of OSS development. And the evolvement is vividly demonstrated in our social network analysis.

In a way, the social structure of OSS projects is both centralized and decentralized. On one hand, it is centralized in the sense that there is a core that consists of key members. These key members are responsible for various issues encountered during the development process. On the other hand, it is decentralized in the sense that the decision or communication core is not concentrated on one or two members but a group of key members.

Like any other research, this research has its share of limitations. First, the cases were only selected from SourceForge.net. Although SourceForge is the world’s largest Web site hosting open source software, there are also some other similar Web sites. Therefore, the total number of OSS projects in SourceForge cannot be viewed as the whole population. However, as we argued before, SourceForge is probably the best data collection site for this research.

Second, the bug tracking system was chosen as our data resource. The selection of bug tracking system as our research setting and data resource may have had an effect on the outcome and results. Besides the bug tracking forum, there are other forums that also provide space for participants to communicate with one another, such as mailing lists and feature requests. However, as we highlighted earlier, the bug systems are the most active forum, providing rich interaction data. The bug tracking systems also represent the spirit of open source software development. Examining the interaction data from other forums can be one of our research extensions in the future.

Third, because our research objective is to investigate interaction pattern, we chose projects that have a relatively large number of developers, a large number of bug reports, and relatively long history. Although we tried to involve different types of projects (i.e., different project sizes, project types, and intended audience), these three cases may not be representatives of OSS projects, for example, small projects with only one or two developers and few interactions. Increasing the sample size and including various types of OSS projects is one of our future research directions.
Social Network Structures in Open Source Software Development Teams

IMPLICATIONS AND CONCLUSION

This paper examines the interaction patterns of OSS teams. The research findings suggest that the interaction structure starts from a single hub and evolves to a core/periphery model. We argue that the social structure of OSS teams is both centralized and decentralized. It is centralized in the sense that there exists a relatively stable core that consists of a group of key developers. It is also decentralized because of distributed decision making among key developers and the broad collaboration between developers and users as well as among developers themselves.

The paper presents the evolution of the social structure of OSS projects from a longitudinal perspective. It also provides empirical evidence of the change of interaction patterns from a single hub to a core/periphery model. Moreover, the paper utilizes social network analysis as the research method. This approach has been shown in this research as an effective tool in analyzing the social structure in OSS teams.

Social structure is an important variable for understanding social phenomenon. Open source software, with its open and unique nature, attracts researchers to ask a series of questions. For example, how do participants of OSS projects interact and collaborate with each other? What factors facilitate the interaction and the collaboration? And further, how does the collaboration affect project performance of OSS teams? Social network analysis is a good approach to investigate these questions. This research represents a pioneering effort in this direction.

REFERENCES


Social Network Structures in Open Source Software Development Teams


ENDNOTE

1 The Web address for SourceForge is www.sourceforge.net.


Chapter 2.18
A Graphical User Interface (GUI) Testing Methodology

Zafar Singhera
ZAF Consulting, USA

Ellis Horowitz
University of Southern California, USA

Abad Shah
R & D Center of Computer Science, Pakistan

ABSTRACT

Software testing in general and graphical user interface (GUI) testing in particular is one of the major challenges in the lifecycle of any software system. GUI testing is inherently more difficult than the traditional and command-line interface testing. Some of the factors that make GUI testing different from the traditional software testing and significantly more difficult are: a large number of objects, different look and feel of objects, many parameters associated with each object, progressive disclosure, complex inputs from multiple sources, and graphical outputs. The existing testing techniques for the creation and management of test suites need to be adapted/enhanced for GUIs, and new testing techniques are desired to make the creation and management of test suites more efficient and effective. In this article, a methodology is proposed to create test suites for a GUI. The proposed methodology organizes the testing activity into various levels. The tests created at a particular level can be reused at higher levels. This methodology extends the notion of modularity and reusability to the testing phase. The organization and management of the created test suites resembles closely to the structure of the GUI under test.

INTRODUCTION

Graphical user interfaces (GUI) are an important part of any end-user software application today and can consume significant design, development, and testing activities. As much as half of the source code of a typical user-interaction intensive application can be related to user inter-
A Graphical User Interface (GUI) Testing Methodology

faces (Harold, Gupta, & Soffa, 1993; Horowitz & Singhera, 1993). GUIs provide an easier way of using various functions of the application by organizing them in a hierarchy of options and presenting only the options which make sense in the current working context. GUIs help users concentrate on the problem instead of putting efforts in remembering all the options provided by the software application that is being used to solve the problem, or searching for the right option from a huge list of options provided by the application. Graphical user interfaces organize the standard user actions and working paradigms into various components that are presented graphically to the user during various usage and application contexts. GUIs enhance the usability of an application significantly. However it also makes application development, testing and maintenance significantly more difficult (Myers, 1993; Wittel & Lewis, 1991). The nature of GUI applications, their asynchronous mode of operation, nontraditional input and output, and hierarchical structure for user interaction make their testing significantly different and difficult from the traditional software testing.

Functional and regression testing of graphical user interfaces is significantly more complex than testing of traditional non-GUI applications because of the additional complexities mentioned in the previous paragraph. A number of commercial tools, like Mercury Interactive’s WinRunner, XRunner and Segue Software’s SilkPerformer, are used in the industry to test graphical user interfaces. These tools provide capture/replay capabilities to test a graphical user interface. Although functionality provided by these tools is sufficient for typical recorded/replay scenarios but they lack an underlying model that can provide more information about the test coverage or to determine the quality of the user interface from a particular functional or implementation perspective. These tools also do not provide a framework that assists in organized and modular testing. The methodology presented in this article uses user interface graphs (UIG) as a framework for organization of test scripts, generation of modular test suites, and coverage analysis of a test execution.

In this article, we propose a methodology for regression testing of graphical user interfaces, with and without a formal specification of the application under test. The remainder of this article is organized as follows: Section 2 highlights some of the best practices and recommendations that help in testing a GUI application in an organized fashion, improve efficiency and effectiveness of testing, reduces possibility of errors, and minimizes repeated work. Section 3 describes the major steps of the proposed methodology. It also introduces a sample X application, called Xman, which is used to demonstrate the effectiveness of the suggested strategy. Section 4 demonstrates the testing methodology when formal specifications of the application under test are not available. Section 5 illustrates the proposed testing methodology when formal specifications of the application under test are available. This section also describes the way statistics are collected during a testing activity and how those can be used to improve the quality of the testing. Section 6 points out the situations when a modification to the application under test might require tuning or recapturing of some of the test scripts. Section 7 concludes the article by summarizing our contribution and providing hints about the future related work.

GUI TESTING: BEST PRACTICES AND RECOMMENDATIONS

In this section, we highlight some of the sought features, well-known best practices and recommendations for planning a testing activity for a graphical user interface.
A Graphical User Interface (GUI) Testing Methodology

- Every element of the GUI should be considered as an object uniquely identifiable by a name. The objects should have a well-defined set of parameters and their response to the outside events should also be well defined.
- The testing activity should be planned carefully around a formal model of the application under test. This model should be powerful enough to provide automatic test generation and coverage analysis.
- Testing of a GUI should be performed in a layered fashion. The list of objects to be tested at a particular level, is either built dynamically while testing at lower levels or from the specifications of the application under test. The list of objects for the lowest level is the basic widget, supported by the underlying toolkit. While testing the highest level, it considers the entire application as a single object. The decision about the number of testing levels and the qualifying criteria for a particular testing level must be made before creating any tests.
- The tests should be organized as a hierarchy of scripts, that is, files containing commands to simulate user actions and verify results. This hierarchy should closely correspond to the object hierarchy of the application under test. Each directory in the hierarchy holds scripts that are related to a particular object and its descendents. The individual scripts should be as small as possible and should test one particular feature of each object. However, if the features are related and simple, then they can be grouped together in the same script.
- Each script should begin with a clean and precise description of the intended purpose of the script and the state of the application required for its proper execution. A script should be divided into three sections. The first section of the script builds the environment required to test the particular feature of the application, the script is intended for. The second section of the script tests the intended feature of the application being tested by the script. The third section restores the state of the AUT and the operating environment, to a point that existed before entering the script.
- A script should be created in such a way that some or all the sections of the script can be executed by calling the script from another script. It provides reusability feature in testing also.
- Instead of manually capturing or replaying the test scripts, a tool should be used to perform these functions automatically and verify the behavior of the application under test. The tool should be capable of addressing an object in the GUI by its symbolic name, instead of its location, dimensions or any other contextual information.
- The data for the result verification should be captured in terms of object attributes when possible and only those attributes should be captured which are critical to verify the functions of the application, being tested by the current script. If image comparisons are unavoidable, then the images should be captured with reference to the smallest enclosing object and area of the captured images should not be more than absolutely required. The number of verifications should also be kept to an absolute minimum especially when image comparisons are involved.
- The script commands to simulate user actions during the replay and the data for verification of the AUT behavior should be kept separately. This separation is required because the verification data might change depending on the environment while the script commands should be independent of the environment and should be valid across multiple platforms. If script commands and verification data are stored separately, then it is easier to port a test suite across multiple platforms. In fact, a good tool should
automatically perform the porting from one hardware platform to the other.

Proposed Methodology for GUI Testing

This section proposes a methodology for the testing of a graphical user interface (GUI). This proposed methodology is suitable particularly when one has a tool similar to Xtester (Horowitz & Singhera, 1993). It follows the recommendations provided in the previous section. The methodology works in the following two scenarios:

i. Testing without formal specifications/model of the application under test
ii. Testing with a formal model of the application.

Both these scenarios are described in the following subsections.

Testing without a Formal Model

Creation of formal specifications of a GUI application for its testing purposes is a difficult task and requires a significant amount of effort. It is also not feasible to invest resources in creating the formal specifications of the application; hence the testing has to be performed without it. The best thing that can be done in such a situation is to incrementally build a test hierarchy for the application, by capturing-user sessions in an organized way. Automatic test generation or coverage analysis is not possible without formal specifications of the application under test. The major steps of the proposed methodology to test an application without a specification are given below:

Step 1: Initialization

Make basic decisions about the testing activity. Some of the most important decisions, which must be taken at this point, are:

Display the object on the screen and verify its appearance;
If the object has sub-objects
   Add its immediate sub-objects at the top of the list;
If the object qualifies for a higher testing level
   Add it to the list of objects for the higher level;
Send expected events to the object and verify its response;
If a new object appears in response to the event
   if the new object is not listed as a tested object
      Add it to the end of the list;
      Pop down the object from the screen;
Mark the object in the list as tested;
• The number of testing levels and criteria to build list of objects for a particular level.
• Initialize a list of objects for each level that holds the names and some information about the objects of the user interface. The information includes the way the object can be mapped on the screen, the mechanism to unmap it from the screen, and if it has been tested or not.
• The location of the test suite and its organization.
• The application resources that will be used during this testing activity should be listed.

Step 2: Building the Initial Object List

Go through the documentation of the application under test and find all the top-level windows, which might appear as starting windows of the application. These windows and their related information is added to the list of objects for the first testing level and marked as tested.

Step 3: Building Test Suite

Take the first object from the top of the object list, which has not been tested, and create test scripts for it. The procedure for creating test scripts for a particular object is given in Figure 1. The sub-objects of the object under test are added to the list of objects by scanning the object from left to right and top to bottom. Keep on taking objects from the top of the object list and testing them until all the objects in the list are marked as tested. When the list associated with a particular level has no untested object, start testing objects from the list, associated with the next higher level. This process continues until all the levels are tested.

Step 4: Creating Script Drivers

Write higher level scripts for all the top level windows for any other complex objects, each of the testing levels and for the entire test suite. These scripts will replay all the scripts related to the object and its descendents. The highest-level script driver should replay each and every script in the suite.

Step 5: Testing the Test Suite

Make sure that all the scripts and script drivers work properly and cover all the features of the application, which needs to be tested. One cannot do much automatically to determine the quality of a test suite, in the absence of a formal model of the application under test. After the creation of the test suite, run the highest-level script driver to verify that all the scripts in the suite are capable of properly replaying and trying to match the features covered by the test suite with those included in the test requirements or application documentation.

Testing with a Formal Model

The strategy to build test scripts without a formal specification, discussed in the previous subsection, puts a lot of responsibility on the person creating those scripts. The strategy also requires that the application under test should be running reliably before the capturing of script is even started. The scripts created without any formal specification are also vulnerable to any modification in the application, which affects its window hierarchy. It requires that after making any changes to the application under test, the affected scripts should be located manually and recaptured or tuned to offset the modification in the application. It is also not possible to create test scripts automatically or to get a coverage measure after running a set of test suites. To overcome these drawbacks and get access to advanced features like automatic test generation and coverage analysis, one has to invest
some effort to formally specify the application under test. This section provides a methodology to test an application when its formal specification is provided or resources are available to build such a specification. The following are the major steps of the testing methodology when a formal specification of an application under test is available.

**Step 1: Building the Model**

Build a user interface graph of an application under test. When resources permit, the very first step in testing the application should be to build a formal model of the application under test. XTester provides such a formal model, called user interface graph (UIG). UIG provides information about the object hierarchy of the application. It also provides information about the nature of a particular object and the effects of an event in an object to the other objects in the user interface. A UIG can be built manually by creating a user interface description language (UIDL) file, or it can be created semi-automatically by steering through the application under test and filling in the missing information about the objects (see Horowitz & Singhera, 1993) for more details on syntax of UIDL and UIG).

**Step 2: Initialization**

Make basic decisions about the testing activity. Some of the most important decisions, which must be taken at this point, are listed as follows:

- The number of testing levels and qualifying criteria for each testing level.
- The location of the test suite and its organization.
- The application resources that will be used during this testing activity.

**Step 3: Build Object Lists**

Build a list of objects for each testing level. After building the formal model, it is possible to build the object lists for all testing levels. The procedure for building those lists is to start from the root of the UIG and perform a post-order walk of the object hierarchy. Add each visited node to the object lists associated with the levels, for which it qualifies.

**Step 4: Building Test Suite**

The strategy for capturing scripts without any formal specification, which has been discussed in the previous section, can also be used for capturing scripts when the application has been specified formally. However, capturing scripts with formal specifications provides us some additional advantages over the scripts which have been captured without any specification. These advantages include an overall picture of the application under test, and hence a more efficient test suite, a test suite which is less affected by the changes in the application, automatic test generation, and coverage analysis.

**Step 5: Creating Script Drivers**

Write higher level scripts for all the top level windows or any other complex objects, each of the testing levels and for the entire test suite. These scripts will replay all the scripts related to the object and its descendents. A highest-level script driver should replay each and every script in the suite. These scripts can also be created automatically.

**Step 6: Coverage Analysis**

Once a test suite has been created; it should be replayed in its entirety to determine the coverage provided by the test suite. This coverage should
be performed at each level, that is, the coverage criteria for level-I should be the verification of all the objects in the application that have been created, mapped, unmapped and destroyed, at least once and every event expected by an object, has been exercised on it at least once. The criteria for higher levels is to make sure that all the interactions and side effects among objects which make a composite object at the corresponding level, has been verified.

**AN INTRODUCTION TO XMAN**

In this section, we introduce the application Xman, which is used to demonstrate the effectiveness of the methodology, discussed in the previous section. Xman is a small application which is distributed with the standard $X$ release. It provides a graphical user interface to the UNIX *man* utility. It has been developed using the Athena widget set and some of its windows are shown in Figure 1. The following paragraphs briefly describe the functionality of Xman.

When Xman is started, it displays its main window, called Xman, by default. This main window contains three buttons: Help, Manual Page and Quit. Clicking on the manual page button, it displays a window, called manual page. A Help window is displayed when the help button is clicked. The quit button is used to exit from Xman.

The manual page window is organized into various sections. A breathe top of the window contains two menu buttons, options and sections, in the left half and a message area in the right. The rest of the area below the bar, called text area, is used to display the names of the available manual pages in the currently selected section, and the contents of the currently selected manual page. Both the names and contents portions of the text area are resizable and have vertical scrollbars on the left.

The Options menu contains the following entries:

- **Display Directory**: It displays names of manual pages in the entire text area.
- **Display Manual Page**: It displays contents of the currently selected manual page in the entire text area.
- **Help**: It displays a help window.
- **Search**: It displays a dialog box to enter the name of a manual page to search for.
- To show both screens, the area is vertically divided into two halves, with the upper half showing the directory contents of currently selected man page section and the lower half showing the contents of the currently selected manual page. This option toggles to Show One Screen and also disables menu entries Display Directory and Display Manual Page.
- **Remove This Manpage**: It removes the Manual Page window from the screen.
- **Open New Manpage**: It creates another Manual Page window.
- **Show Version**: It displays the current version of Xman in the Message area
- **Quit**: exits from the Xman application.

The Sections menu contains one option for each manual page section available on the system. The standard options are User Commands, System Calls, Subroutines, Devices, File Format, Games, Miscellaneous, and System Administration.

The Help window displays a limited version of the Manual Page window. The window has exactly the same structure as the Manual Page window but the Sections menu button and the first five options in the Options menu are disabled. It displays man page for Xman itself. No more than one Help window can exist at a time while an arbitrary number of Manual Page windows can be created.

The testing of Xman has been organized in three layers. The first layer verifies the behavior of
individual GUI objects. The second layer verifies the inter-object effects among objects belonging to the same top level window. The third layer verifies the inter-object effects among objects belonging to different top level windows. The following sections provide details of this testing activity.

TESTING WITHOUT FORMAL SPECIFICATIONS

This section provides a demonstration for the testing of Xman, when no formal model is available for it. The following subsections demonstrate each step of the methodology, described in Section 2.

Initialization

• **Number of testing levels:** As Xman is a fairly small and simple application, so the testing activity is organized in three (3) levels. The first level tests the individual objects in Xman. The second and third levels verify the interactions and side effects of the objects which belong to the same top level window and different top level windows, respectively.

![Figure 2. Main windows of Xman](image-url)
Figure 3: Test Suite Hierarchy of Xman
• **Location of the test suite:** Suppose that the root directory for the test suite being captured, is Xman Test. This directory contains resource file(s) used during testing, and the result files are created in the directory, by default. It also has three subdirectories, that is, Level-1, Level-2 and Level-3, one for each testing level.

• **Object list:** A list of objects, called *Obj List*, is initialized. This list contains information about the objects and is initialized to be empty.

• **Application resources:** ~/XmanTest/Xman.Defaults is the file which contains the default application resources of Xman for this testing activity and these resources always remain the same.

**Building the Initial Object List**

By default, Xman starts with a top box, called Main Window in Figure 2. However, a command line option, -notopbox, is available which can be used to bypass the Main Window and display the Manual Page window directly. As Main Window and Manual Page window are the only windows, which can appear when Xman is started, so the initial object list contains only these two objects.

**Building the Test Suite**

This section provides the details on building all the three testing levels for Xman. To make things simple, we only discuss the creation of test suites related to the Main Window of Xman. We ignore any keyboard accelerators to which Xman responds. The test suites for other windows can be created in a similar way. Figure 3 provides the hierarchy of scripts related to the Main Window for all the three testing levels.

**First Level**

Suppose that the initial object list is built in such a way so that Xman Main Window is on the top of the list. We select it as the object under test and create a new directory, called ~/XmanTest/Level-1/Xman to build test scripts related to the first level testing. The scripts related to Xman Main window object display itself the window on the screen, exercise all the window manager operations on it, and then finally pop-down the window. DspXmanWin.scr script pop-up the Xman window and verifies that looks right. RemoveXmanWin.scr script in its entering section to display the Xman window. It verifies the window manager operations in its core section and then uses RemoveHelpWin.scr script in its leaving section to pop-down the Xman Window. As soon as the Xman window pops up on the screen, we see that it contains four objects, i.e., Manual Browser Label, Help, Quit and Manual Page. There are buttons which are Quit and Manual Page. Manual Browser label is a static piece of text and does not respond to any user actions, so we do not need a script for it. The other three (3) objects are active objects and respond to user events so we create one script for each of them. The entering section of each one of these scripts calls DspXmanWin.scr to display the Xman Main Window on the screen. The ending section of each of these scripts call RemoveXmanwin.scr to remove the Xman Main Window from the screen. The core section of Help Button.scr script verifies the behavior of Help button in Xman Main Window when it is clicked on by a mouse button. The core sections of QuitButton.scr and ManPageButton.scr scripts verify the same thing for Quit and Manual Page buttons in Xman Main Window.
Second Level

The object list of the second level contains all the top level windows of Xman. As we are considering the **Main Window** only in this discussion so we assume that it is at the top of the list and is selected for the testing. There is not much interaction going on in the objects which belong to the Xman **Main Window**. The only interaction is the disappearance of **Main Window**, in response to click on the Quit button. So there will be only one script related to the Main Window which will verify that a click on the Quit button actually destroys the Main Window of Xman. This script is called **MainWindow.scr** and is located in ~*/XmanTest/Level2/*/. This script is also used DspXmanWin.scr and RemoveXman.scr script to display and remove the Main Window from the screen. Another potential script, let us call it XmanOrder.scr, related to the Main Window verifies that the order in which Help or Manual Page buttons are pressed is insignificant. No matter the Help button is pressed before or after the Manual Page button, it displays the Help window properly. The same is also true for the Manual Page button.

Third Level

The object list of the third level includes the root object only, and tests any interactions among the top level windows of Xman can be done. Such interactions which involve the Main Window of Xman include display of the Help window and the Manual Page window in response to mouse clicks on the Help and the Manual page buttons, respectively. Similarly, it also includes disappearance of all the windows related to Xman in response to a click on the Quit button. The three scripts provided at this level, that is, Help.scr, ManualPage.scr and Quit.scr, verify the behavior, related to the corresponding button, mentioned above. This level might also include scripts which verify application behavior, like multiple clicks on the Help button and do not create more than one Help windows while each click on the Manual Page button create a new Manual Page window.

Creating Script Drivers

Once all the scripts for Xman has been captured, we need driver scripts so that all the scripts in the entire suite, all the scripts in a particular testing level or all the scripts related to a particular object can be executed automatically in the desired sequence. For example, we create a script driver, at each testing level, which executes all the scripts created for testing Xman Main Window and its descendents, at that particular level. These scripts are ~*/XmanTest/Level-1/Xman/All.scr, ~*/XmanTest/Level-2/Xman/All.scr, and ~*/XmanTest/Level-3/Xman/All.scr, respectively. The script ~*/XmanTest/Level-1/All.scr drive all the scripts created for the first testing level and similarly the other two drivers execute scripts related to the other two levels. The script ~*/Xman-Test/All.scr drives all the scripts in all the three levels of the test suite.

Testing the Test Suite

After the creation of the test suite, it is necessary to replay all the scripts in the suite and verify if they work properly and also to make sure that they cover all the features which need to be tested. Although, without a formal specification, it is impossible to perform any reasonable automatic coverage analysis but at least the replayed events and the objects which appear during the replay, can be matched against application documentation to determine if any object or event has not been covered by the generated test suite.

TESTING WITH FORMAL SPECIFICATIONS

This section demonstrates the testing of Xman when we have enough resources to build a formal
model for Xman. The following subsection illustrates each step of the methodology, described in Section 2.2.

**Building the Model**

When the resources permit, the first step for testing is building a formal model for the application under test. Figure 4 displays a User Interface Graph built for Xman. The Root node of the graph represents the root window of the screen. The children of the Root node represent the six top level windows of Xman. The nodes at lower levels in the graph represent the descendants of the top level windows. Let us take the main window of Xman as an example. It is represented as MainWin node in the User Interface Graph. The child of the MainWin node is the Form node which acts as a container widget for the buttons and the label in the main window. The ManPage, Quit and Help nodes represent the Manual. Page, Quit and Help command buttons in the main window, respectively. The label node represents the Manual Browser label in the main window. The dark black arc from the ManPage node to the Manual Page node represents the fact that clicking a mouse button over the ManPage button in the main window affects the top level window, called Manual Page. The arc from the Quit node to the Root node represents that a click on the Quit button affects all the top level windows of Xman. The type of event represented by an arc is reflected by the drawing pattern of the arc. The
two arcs mentioned above have the same pattern and represent button clicks. An arc with a different pattern is the arc from the Text node to the search node. This pattern represents text entry and the arc represents that entering text in the Text node affect the search dialog box.

**Initialization**

All the decision and actions taken at the initialization step, that is, the number of testing levels and their organization, the location of the test suite and the application default resources, is kept the same as for testing without a formal specification, described in Section 4.1.

**Building Object Lists**

After building the user interface graph for Xman, it is possible to build object lists for all levels of testing. This can be done either manually or automatically by scanning each and every node in the User Interface Graph and verifying if it qualifies to be tested on a particular testing level. All the objects in Xman qualify for the first testing level and hence are placed on the list associated with it. The qualifying criterion for the second level is that the object must be a top level window, that is, its corresponding node must be a child of the root of the user interface graph. Some of the objects which qualify for the second testing level are Xman Main Window, Manual Page window, Help window, Search dialog box and so forth. The third level treats the entire Xman application as a single object, and its corresponding node, the root, is the only object which qualifies for the third level of testing.

**Building Test Suite**

The strategy for capturing scripts without any formal specifications, discussed in Section 4.3, can be used for capturing scripts when the application has been specified formally. However, capturing scripts with formal specifications also provides us the capability to write the test scripts manually or generate test scripts automatically by using the formal model and the test requirements. All of these three techniques for building a test suite are explained in the following section.

**Capturing Scripts**

A tool can be used for capturing user sessions for building a test suite, in exactly the same way as for capturing without a formal model, as mentioned in Section 4.3. However, the presence of a formal model makes the generated scripts more robust and easy to follow and debug. A formal model also provides the flexibility that the captured scripts can be executed in any order without any conflict in window names. It becomes easier to modify the test suite in case of a modification to the application under test. In fact in most cases the modification can be reflected in the specification file and the test suite remains valid without making any changes.

**Writing Scripts Manually**

The formal specifications of the application, under test, also allows the user to write scripts manually, even for the first level of testing. This feature is particularly helpful when the test scripts need to be created in parallel with the application development in order to reduce the total development time. The way it can be organized is to formally specify the graphical user interface of the application once the design is complete. The developers and testers agree on this formal specification and any future changes are properly communicated between the two groups. Having the formal specification of the graphical user interface at hand, the testers can develop the scripts manually in parallel with the application development. Once the object hierarchy and the behavior of the objects is known, the manual writing of test scripts is as easy as writing a UNIX shell script.
Automatic Test Generation

The formal specifications of an application under test also provide capabilities to generate test scripts automatically. A tool can be developed, which will read the specifications of an application and create test scripts for a particular level of testing or the entire test suite. Similar tools can be developed and used to generate test scripts to test all the objects of a particular type or all the user actions of a particular type. For example, such a tool can generate test scripts to exercise clicks on each command button in an entire application. A similar tool can generate a suite of test scripts to verify that all windows in an application under test are displayed on the screen at least once. Another tool might generate scripts to individually select all the options of each menu in the application and verify that the application responds in the expected manner. There can be another tool that will create scripts for the selection of multiple menu options and/or command buttons in different sequences to verify the application response. All the mentioned tools will only create scripts and validation data that have to be captured by replaying these scripts and by using caprep (Horowitz & Singhera, 1993) or a similar tool to replay these scripts in Update mode.

The logic behind these automatic test generation tools is the same as it is used in Section 5

Figure 5. Pseudo Code for Automatic Test Generator

```
Build the User Interface Graph of the application under test;
Build an object list by a pre-order traversal of the User Interface Graph using parent-child relationship arcs.
for each "element on the list"
  do
    if the element is a top level window
      Create a new directory and change to the directory.
      Display the element on the screen.
    fi
    if the element accepts any user events
      Create a script for the element
      for each kind of user event accepted by the element
        do
          Add commands in script to
          Generate the event on the element;
          Verify the effect of that event;
          Undo the effect of the event;
        done
    fi
  done
```
A Graphical User Interface (GUI) Testing Methodology

for manual creation of test suites. The tool starts at the root of the UIG and builds a list of GUI elements by performing a pre-order walk of the UIG. During this walk only the arcs that represent parent-child relationships are considered. After building this list, its entries are taken one by one to create test scripts for them. If the current GUI element, taken from the list, is a top-level window, then a separate directory is created for it and the scripts for the element and all of its

Figure 6. Statistics file created by XTester

<table>
<thead>
<tr>
<th>Object Name</th>
<th>CW</th>
<th>DW</th>
<th>MW</th>
<th>UMWB</th>
<th>BR</th>
<th>KP</th>
<th>KR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xman</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Xman*Form</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Xman*ManualPage</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>50</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>Xman*Quit</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Xman*Help</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Xman*Label</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>StandBy</td>
<td>28</td>
<td>23</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>StandBy*Message</td>
<td>28</td>
<td>23</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LikeToSave</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LikeToSave*Form</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LikeToSave*Message</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LikeToSave*Yes</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LikeToSave*No</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Help</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Help*Form</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Help*MenuBar</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Help*Options</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>20</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Help*Sections</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Help*Message</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Help*TextArea</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Help*Serollbar</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Help. Form. 3</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Manpage</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*Form</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*MenuBar</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*Options</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>92</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*Sections</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>18</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*Message</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*TextArea</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*Serollbar</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>ManPage. Form. 3</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*DirArea</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*DirList</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ManPage*List</td>
<td>27</td>
<td>23</td>
<td>27</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
descendents are created in that directory. If the currently selected element belongs to the category for which a script is required, then following the arcs that represent user actions on the element creates one. Figure 4 shows the pseudo code for such an automatic test generator.

Coverage Analysis

No testing activity is useful unless it provides some coverage measures/analysis. This coverage measure reflects the quality of the testing activity. The UIG provides us with a framework to determine such a coverage. During capture or replay of scripts, XTester keeps track of the user actions and their effects on individual objects. This information is available in a .stt file at the end of the testing activity. Currently, the information captured in .stt about a particular object includes the number of times it was created, mapped, unmapped, and destroyed. It also accumulates the number of times a mouse button or keyboard key was pressed or released over the object. This information helps the user to locate any particular objects in the application which have not been created, destroyed, mapped, unmapped, or received a user event. These statistics can also be used for improving the efficiency of the test suite by removing the repetitive testing of the same characteristics, whenever possible. Figure 6 shows a file created by XTester after replaying a certain test suite for Xman. The legend is displayed at the top of the file to describe acronyms for various actions. The next line after the legend provides the heading for the table. Each line in the table provides statistics about a particular object. For example, the very first line of the table provides statistics about the Main Window of Xman, named Xman. This particular line shows that the object named Xman was created once, never destroyed, mapped twice on the screen and unmapped once, by the corresponding test script. It also shows that the corresponding test suite never exercised a button press/release or key press/release events on the Xman object.

Tools can be developed to extract information from a .stt file. Analysis tools are developed that take a .uidl file to build a UIG of the application under test, and a .stt file to get statistics collected from a particular test run. The tool maps the collected statistics on to the object hierarchy and produces an annotated graph. Queries can be run on this annotated graph to determine the effectiveness of the test suite. For example, one can see how

Figure 7. Pseudo code for quality analyzer

| Read in the required criteria for analysis; |
| Build the User Interface Graph of the application under test; |
| Read in the specified statistics file(s); |
| for each element in the User Interface Graph |
| do |
| If the element qualifies for the given criteria |
| Display the required information in proper format; |
| done |

---

A Graphical User Interface (GUI) Testing Methodology
many nodes and arcs in the object hierarchy have not been exercised by this particular test suite, or see the objects and inter-object arcs to determine objects that did not receive an expected user action. Similarly, one can filter out the objects that did not receive an expected user event to create an annotated graph that satisfies a particular criterion.

INVALIDATION OF TEST DATA

XTester captures information as two entities, script commands and verification data and saves them in different files, .scr and .img, respectively. This section describes the scenario in which created test script(s) might fail and has to be re-captured or re-written. The following scenarios invalidate some of the captured scripts:

- An application is modified in such a way that WM_NAME property of a top-level window is changed. This modification will only affect the scripts that have been captured without a specification and are related to the window whose WM_NAME property was changed. The scripts captured with a formal specification remain unaffected, provided the relevant specifications are also modified to reflect the change in the application.

- The application is changed so that the order or number of children of a particular node is changed. The scripts that are captured without a specification and address objects in the modified hierarchy, are affected and have to be recaptured or tuned. However, the scripts captured with a formal specification remain unaffected provided the specification is also modified accordingly. XTester provides option to build either full or short object names. If the application is modified so that the depth of a hierarchy is changed, then all the fully qualified names belonging to that hierarchy will no longer be valid names and has to be modified to reflect the change. However, short names that are relative to their immediate parents will still remain valid.

- If object information has been captured as an image, then trying to replay the scripts on another workstation that is incompatible with the workstation on which image was captured will give false alarms. The scripts work fine across multiple platforms, however, verification data is the platform specific in case of images. An easier way of creating verification data for a new hardware platform will be to replay all the scripts in Update mode that replaces the current verification data with the newly available one.

CONCLUSION AND FUTURE DIRECTIONS

In this article, we have suggested guidelines that are useful in planning a testing activity for a graphical user interface (GUI). We have also presented a methodology for testing a GUI, both when specifications of an application under test is not available, and when such specifications are provided or resource code is available to build such a specification. This article also demonstrates the use of the proposed methodology to test a sample X application, Xman, with or without specifications. It also illustrates how the model is helpful in automatic test generation and coverage analysis. In the end, we describe the situations in which the scripts captured by XTester become invalid.

The methodology and the underlying UIG framework discussed in this article can be very effectively used to model and test web sites and web applications. We are actively working to extend this proposed methodology to be used for testing web-based applications and semantic web applications.
REFERENCES


This work was previously published in International Journal of Information Technology and Web Engineering 3(2), edited by G. Alkhateb and D. Rine, pp. 1-18, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 2.19
Socio–Cognitive Engineering

Mike Sharples
University of Nottingham, Jubilee Campus, UK

INTRODUCTION

Socio-cognitive engineering is a framework for the systematic design of socio-technical systems (people and their interaction with technology), based on the study and analysis of how people think, learn, perceive, work, and interact. The framework has been applied to the design of a broad range of human-centered technologies, including a Writer’s Assistant (Sharples, Goodlet, & Pemberton, 1992), a training system for neuroradiologists (Sharples et al., 2000), and a mobile learning device for children (Sharples, Corlett, & Westmacott, 2002). It has been adopted by the European MOBIlearn project (www.mobilearn.org) to develop mobile technology for learning. It also has been taught to undergraduate and postgraduate students to guide their interactive systems projects. An overview of the framework can be found in Sharples et al. (2002).

BACKGROUND

The approach of socio-cognitive engineering is similar to user-centered design (Norman & Draper, 1986) in that it builds on studies of potential users of the technology and involves them in the design process. But users are not always reliable informants. They may idealize their methods, describing ways in which they would like to or have been told to work, rather than their actual practices. Although users may be able to describe their own styles and strategies of working, they may not be aware of how other people can perform a task differently and possibly more effectively. Surveys of user preferences can result in new technology that is simply an accumulation of features rather than an integrated system.

Thus, socio-cognitive engineering is critical for the reliability for user reports. It extends beyond individual users to form a composite picture of the human knowledge and activity, including cognitive processes and social interactions, styles and strategies of working, and language and patterns of communication. The term actor is used rather than user to indicate that the design may involve people who are stakeholders in the new technology but are not direct users of it.

The framework extends previous work in soft systems (Checkland & Scholes, 1990), socio-technical and cooperative design (Greenbaum & Kyng, 1991; Mumford, 1995; Sachs, 1995), and the application of ethnography to system design.
Socio-Cognitive Engineering

(see Rogers & Bellotti [1997] for a review). It incorporates existing methods of knowledge engineering, task analysis, and object-oriented design, but integrates them into a coherent methodology that places equal emphasis on software, task, knowledge, and organizational engineering.

The framework also clearly distinguishes studying everyday activities using existing technology from studying how the activity changes with proposed technology. It emphasizes the dialectic between people and artefacts; using artefacts changes people’s activities, which, in turn, leads to new needs and opportunities for design.

FRAMEWORK

Figure 1 gives a picture of the flow and main products of the design process. It is in two main parts: a phase of activity analysis to interpret how people work and interact with their current tools and technologies, and a phase of systems design to build and implement new interactive technology. The bridge between the two is the relationship between the Task Model and the Design Concept. Each phase comprises stages of analysis and design that are implemented through specific methods. The framework does not prescribe which methods to use; the choice depends on the type and scale of the project.

It is important to note that the process is not a simple sequence but involves a dialogue between the stages. Earlier decisions and outcomes may need to be revised in order to take account of later findings. When the system is deployed, it will enable and support new activities, requiring another cycle of analysis, revision of the Task Model, and further opportunities for design.

The elements of socio-cognitive engineering are as follows:

- **Project**: The diagram shows the process of design, implementation, and deployment for a single project.
- **Actors**: Different types of people may be involved in or affected by the design and deployment, including (depending on the scale of the project) design, marketing and technical support teams, direct users of the system, and other people affected by it (e.g., administrative staff).
- **Roles**: The actors take on roles (e.g., team leader), which may change during the project.

Figure 1. Overview of the flow and main products of the design process
Socio-Cognitive Engineering

- **Stage:** Each box represents one stage of the project.
- **Methods:** Each stage can be carried out by one or more methods of analysis and design, which need to be specified before starting the stage.
- **Tools:** Each method has associated tools (for activity analysis, software specification, systems design, and evaluation) in order to carry out the method.
- **Outcomes:** Each stage has outcomes that must be documented, and these are used to inform and validate the system design.
- **Measures:** Each design decision must be validated by reference to outcomes from one of the stages.

The general sequence for socio-cognitive engineering is as follows:

1. Form a project team.
2. Produce General Requirements for the project.
3. Decide which methods and tools will be used for each stage of the project.
4. Decide how the process and outcomes will be documented.
5. Decide how the project will be evaluated.
6. Carry out each stage of the project, ensuring that the requirements match the design.
7. Carry out a continuous process of documentation and evaluation.

The process starts by specifying the **General Requirements** for the system to be designed. These provide broad yet precise initial requirements and constraints for the proposed system in language that designers and customers can understand. They are used to guide the design and to provide a reference for validation of the system. The requirements normally should indicate:

- The scope of the project;
- The main actors involved in designing, deploying, using, and maintaining the system;
- The market need and business case; and
- General attributes and constraints of the proposed system (i.e., whether it aims to support individual or collaborative working).

The requirements will be extended and made more precise as the project progresses.

This leads to two parallel studies: a theory-based study of the underlying cognitive processes and social activities, and an investigation into how everyday activities are performed in their normal contexts. The **Theory of Use** involves an analysis of relevant literature from cognitive psychology, social sciences, and business management to form a rich picture of the human knowledge and activity. It is essential that this should offer a clear guide to system design. Thus, it must be relevant to the intended use of the system and extend the requirements in a form that can be interpreted by software designers and engineers.

The aim of carrying out **Field Studies** is to uncover how people interact with current technology in their normal contexts. The role of the fieldworker is both to interpret activity and to assist technology design and organizational change. This addresses the widely recognized problem of ethnographic approaches that, while they can provide an understanding of current work practices, are not intended to explore the consequences of socio-technical change.

Table 1 shows a multi-level structure for field studies, with level 1 consisting of a survey of the existing organizational structures and schedules, levels 2 and 3 providing an analysis of situated practices and interactions of those for whom the technology is intended, and level 4 offering a synthesis of the findings in terms of designs for new socio-technical systems. The four levels give an overview of activity, leading to a more detailed investigation of particular problem areas, with each level illuminating the situated practices and also providing a set of issues to be addressed for the next level. These piece together into a composite picture of how people interact with technology.
in their everyday lives, the limitations of existing practices, and ways in which they could be improved by new technology.

The outcomes of these two studies are synthesized into a Task Model. This is a synthesis of theory and practice related to how people perform relevant activities with their existing technologies. It is the least intuitive aspect of socio-cognitive engineering; it is tempting to reduce it to a set of bullet-point issues, yet it provides a foundation for the systems design. It could indicate:

- The main actors and their activity systems;
- How the actors employ tools and resources to mediate their interaction and to externalize cognition;
- How the actors represent knowledge to themselves and to others;
- The methods and techniques that the actors employ, including differences in approach and strategy;
- The contexts in which the activities occur;
- The implicit conventions and constraints that influence the activity; and
- The actors’ conceptions of their work, including sources of difficulty and breakdown in activity and their attitudes toward the introduction of new technology.

The Design Concept needs to be developed in relation to the Task Model. It should indicate how the activities identified by the Task Model could be transformed or enhanced with the new technology. It should:

- Indicate how limitations from the Task Model will be addressed by new technology;

---

**Table 1. Multi-level structure for field studies**

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Activity structures and schedules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity:</td>
<td>Study work plans, organizational structures, syllabuses, resources.</td>
</tr>
<tr>
<td>Purpose:</td>
<td>To discover how the activities are supposed to be conducted.</td>
</tr>
<tr>
<td>Outcome:</td>
<td>Description of the existing organizational and workplace structures; identification of significant events.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 2</th>
<th>Significant events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity:</td>
<td>Observe representative formal and informal meetings and forms of communication.</td>
</tr>
<tr>
<td>Purpose:</td>
<td>To discover how activities, communication, and social interaction are conducted in practice.</td>
</tr>
<tr>
<td>Outcome:</td>
<td>A description and analysis of events that might be important to system design; identification of mismatches between how activity has been scheduled and how it is has been observed to happen.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 3</th>
<th>Conceptions and conflicts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity:</td>
<td>Conduct interviews with participants to discuss areas of activity needing support, breakdowns, issues, differences in conception.</td>
</tr>
<tr>
<td>Purpose:</td>
<td>To determine people’s differing conceptions of their activity; uncover issues of concern in relation to new technology; explore mismatches between what is perceived to happen and what has been observed.</td>
</tr>
<tr>
<td>Outcome:</td>
<td>Issues in everyday life and interactions with existing technology that could be addressed by new technology and working practices.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 4</th>
<th>Determining designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity:</td>
<td>Elicitation of requirements; design space mapping; formative evaluation of prototypes.</td>
</tr>
<tr>
<td>Purpose:</td>
<td>To develop new system designs.</td>
</tr>
<tr>
<td>Outcome:</td>
<td>Prototype technologies and recommendations for deployment.</td>
</tr>
</tbody>
</table>
Socio-Cognitive Engineering

Table 2. A building-block framework for socio-cognitive system design

<table>
<thead>
<tr>
<th></th>
<th>Software Engineering</th>
<th>Task Engineering</th>
<th>Knowledge Engineering</th>
<th>Organizational Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maintain</strong></td>
<td>Installed system</td>
<td>New task structure</td>
<td>Augmented knowledge</td>
<td>New organizational</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>structure</td>
</tr>
<tr>
<td><strong>Evaluate</strong></td>
<td>Debugging</td>
<td>Usability</td>
<td>Conceptual change, skill</td>
<td>Organizational change</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>development</td>
<td></td>
</tr>
<tr>
<td><strong>Integrate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Implement</strong></td>
<td>Prototypes,</td>
<td>Interfaces,</td>
<td>Knowledge</td>
<td>Communications,</td>
</tr>
<tr>
<td></td>
<td>Documentation</td>
<td>Cognitive tools</td>
<td>representation</td>
<td>Network resources</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td>Algorithms and</td>
<td>Human-computer</td>
<td>Domain map, user</td>
<td>Socio-technical system</td>
</tr>
<tr>
<td></td>
<td>heuristics</td>
<td>interaction</td>
<td>model</td>
<td></td>
</tr>
<tr>
<td><strong>Interpret</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Analyze</strong></td>
<td>Requirements</td>
<td>Tasks: goals,</td>
<td>Knowledge: concepts,</td>
<td>Workplace: practices,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>objects,</td>
<td>skills</td>
<td>interactions</td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td>Existing systems</td>
<td>Conventional task</td>
<td>Domain knowledge</td>
<td>Organizational structures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>structures and</td>
<td></td>
<td>and schedules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>processes</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Propose</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Outline a system image (Norman, 1986) for the new technology;
- Show the look and feel of the proposed technology;
- Indicate the contexts of use of the enhanced activity and technology; and
- Propose any further requirements that have been produced as a result of constructing the Design Concept.

The Design Concept should result in a set of detailed design requirements and options that can be explored through the design space.

The relationship between the Task Model and Design Concept provides the bridge to a cycle of iterative design that includes:

- Generating a space of possible system designs, systematically exploring design option and justifying design decisions;
- Specifying the functional and non-functional aspects of the system;
- Implementing the system; and
- Deploying and maintaining the system.

Although these stages are based on a conventional process of interactive systems design (see Preece, Rogers, & Sharp [2002] for an overview), they give equal emphasis to cognitive and organizational factors as well as to task and software specifications. The stages shown in Figure 1 are an aid to project planning but are not sufficiently detailed to show all the design activities. Nor does the figure make clear that to construct a successful integrated system requires the designers to integrate software engineering with design for human cognition, social interaction, and organizational management. The “building-block diagram in Table 2 gives a more detailed picture of the system’s design process.

The four “pillars” indicate the main processes of software, task, knowledge, and organizational engineering. Each “brick” in the diagram shows
one outcome of a design stage, but it is not necessary to build systematically from the bottom up. A design team may work on one pillar (e.g., knowledge engineering) up to the stage of system requirements, or it may develop an early prototype based on a detailed task analysis but without a systematic approach to software engineering. How each activity is carried out depends on the particular application domain, actors, and contexts of use.

The design activities are modular, allowing the designer to select one or more methods of conducting the activity according to the problem and domain. For example, the usability evaluation could include an appropriate selection of general methods for assessing usability, or it could include an evaluation designed for the particular domain.

It should be emphasized that the blocks are not fixed entities. As each level of the system is developed and deployed, it will affect the levels that follow (e.g., building a prototype system may lead to revising the documentation or re-evaluating the human-computer interaction; deploying the system will create new activities). These changes need to be analyzed and supported through a combination of new technology and new work practices. Thus, the building blocks must be revisited both individually to analyze and update the technology in use, and through a larger process of iterative redesign.

Although Table 1 shows system evaluation as a distinct phase, there also will be a continual process of testing to verify and validate the design, as shown in Figure 1. Testing is an integral part of the entire design process, and it is important to see it as a lifecycle process (Meek & Sharples, 2001) with the results of testing early designs and prototypes being passed forward to provide an understanding of how to deploy and implement the system, and the outcomes of user trials being fed back to assist in fixing bugs and improving the design choices.

The result of the socio-cognitive engineering process is a new socio-technical system consisting of new technology, its associated documentation, and proposed methods of use. When this is deployed in the workplace, home, or other location, it not only should produce bugs and limitations that need to be addressed but also engender new patterns of work, social, and organizational structures that become contexts for further analysis and design.

FUTURE TRENDS

The computer and communications industries are starting to recognize the importance of adopting a human-centered approach to the design of new socio-technical systems. They are merging their existing engineering, business, industrial design, and marketing methods into an integrated process, underpinned by rigorous techniques to capture requirements, define goals, predict costs, plan activities, specify designs, and evaluate outcomes. IBM, for example, has developed the method of User Engineering to design for the total user experience (IBM, 2004). As Web-based technology becomes embedded into everyday life, it increasingly will be important to understand and design distributed systems for which there are no clear boundaries between people and technology.

CONCLUSION

Socio-cognitive engineering forms part of an historic progression from user-centered design and soft systems analysis toward a comprehensive and rigorous process of socio-technical systems design and evaluation. It has been applied through a broad range of projects for innovative human technology and is still being developed, most recently as part of the European MOBIlearn project.
REFERENCES


KEY TERMS

**Activity System**: The assembly and interaction of people and artefacts considered as a holistic system that performs purposeful activities. See http://www.edu.helsinki.fi/activity/pages/chatanddwr/activitysystem/

**Human-Centred Design**: The process of designing socio-technical systems (people in interaction with technology) based on an analysis of how people think, learn, perceive, work, and interact.

**Socio-Technical System**: A system comprising people and their interactions with technology (e.g., the World Wide Web).

**Soft Systems Methodology**: An approach developed by Peter Checkland to analyze complex problem situations containing social, organizational, and political activities.

**System Image**: A term coined by Don Norman (1986) to describe the guiding metaphor or model of the system that a designer presents to users (e.g., the desktop metaphor or the telephone as a “speaking tube”). The designer should aim to create a system image that is consistent and
familiar (where possible) and enables the user to make productive analogies.

**Task Analysis:** An analysis of the actions and/or knowledge and thinking that a user performs to achieve a task. See [http://www.usabilitynet.org/tools/taskanalysis.htm](http://www.usabilitynet.org/tools/taskanalysis.htm)

**User-Centered Design:** A well-established process of designing technology that meets users’ expectations or that involves potential users in the design process.

**User Engineering:** A phrase used by IBM to describe an integrated process of developing products that satisfy and delight users.
Chapter 2.20
On the Cognitive Processes of Human Perception with Emotions, Motivations, and Attitudes

Yingxu Wang
University of Calgary, Canada

ABSTRACT

An interactive motivation-attitude theory is developed based on the Layered Reference Model of the Brain (LRMB) and the object-attribute-relation (OAR) model. This paper presents a rigorous model of human perceptual processes such as emotions, motivations, and attitudes. A set of mathematical models and formal cognitive processes of perception is developed. Interactions and relationships between motivation and attitude are formally described in real-time process algebra (RTPA). Applications of the mathematical models of motivations and attitudes in software engineering are demonstrated. This work is a part of the formalization of LRMB, which provides a comprehensive model for explaining the fundamental cognitive processes of the brain and their interactions. This work demonstrates that the complicated human emotional and perceptual phenomena can be rigorously modeled and formally treated based on cognitive informatics theories and denotational mathematics.

INTRODUCTION

A variety of life functions and cognitive processes has been identified in cognitive informatics (Wang, 2002a, 2003a, 2003b, 2007b) and cognitive psychology (Payne & Wenger, 1998; Pinel, 1997; Smith, 1993; Westen, 1999; Wilson & Keil, 1999). In order to formally and rigorously describe a comprehensive and coherent set of mental processes and their relationships, an LRMB has been developed (Wang & Wang, 2006; Wang, Wang, Patel, & Patel, 2006) that explains the functional mechanisms and cognitive processes of the brain and the natural intelligence. LRMB encompasses 39 cognitive processes at six layers known as the sensation, memory, perception, action, meta and higher cognitive layers from the bottom up.
Definition 1: Perception is a set of internal sensational cognitive processes of the brain at the subconscious cognitive function layer that detects, relates, interprets, and searches internal cognitive information in the mind.

Perception may be considered as the sixth sense of human beings since almost all cognitive life functions rely on it. Perception is also an important cognitive function at the subconscious layers that determines personality. In other words, personality is a faculty of all subconscious life functions and experience cumulated via conscious life functions. It is recognized that a crucial component of the future generation computers known as the cognitive computers is the perceptual engine that mimic the natural intelligence (Wang, 2006a, 2007c).

The main cognitive processes at the perception layer of LRMB are emotion, motivation, and attitude (Wang et al., 2006). This article presents a formal treatment of the three perceptual processes, their interrelationships, and interactions. It demonstrates that complicated psychological and cognitive mental processes may be formally modeled and rigorously described. Mathematical models of the psychological and cognitive processes of emotions, motivations, and attitudes are developed in the following three sections. Then, interactions and relationships between emotions, motivations, and attitudes are analyzed. Based on the integrated models of the three perception processes, the formal description of the cognitive processes of motivations and attitudes will be presented using RTPA (Wang, 2002b, 2003c, 2006b, 2007a). Applications of the formal models of emotions, motivations, and attitudes will be demonstrated in a case study on maximizing strengths of individual motivations in software engineering.

THE HIERARCHICAL MODEL OF EMOTIONS

Emotions are a set of states or results of perception that interprets the feelings of human beings on external stimuli or events in the binary categories of pleasant or unpleasant.

Definition 2: An emotion is a personal feeling derived from one’s current internal status, mood, circumstances, historical context, and external stimuli.

Emotions are closely related to desires and willingness. A desire is a personal feeling or willingness to possess an object, to conduct an interaction with the external world, or to prepare for an event to happen. A willingness is the faculty of conscious, deliberate, and voluntary choice of actions.

According to the study by Fischer, Shaver, and Carnochan (1990) and Wilson and Keil (1999), the taxonomy of emotions can be described at three levels known as the sub-category, basic, and super levels as shown in Table 1.

It is interesting that human emotions at the perceptual layer may be classified into only two opposite categories: pleasant and unpleasant.

Table 1. Taxonomy of emotions

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Super level</td>
<td>Positive (pleasant)</td>
</tr>
<tr>
<td></td>
<td>Negative (unpleasant)</td>
</tr>
<tr>
<td>Basic level</td>
<td>Bliss, pride, contentment</td>
</tr>
<tr>
<td></td>
<td>Love</td>
</tr>
<tr>
<td></td>
<td>Joy</td>
</tr>
<tr>
<td></td>
<td>Anger</td>
</tr>
<tr>
<td></td>
<td>Sadness</td>
</tr>
<tr>
<td></td>
<td>Fear</td>
</tr>
<tr>
<td>Sub-category level</td>
<td>Fondness, infatuation</td>
</tr>
<tr>
<td></td>
<td>Annoyance, hostility, contempt, jealousy</td>
</tr>
<tr>
<td></td>
<td>Agony, grief, guilt, loneliness</td>
</tr>
<tr>
<td></td>
<td>Horror, worry</td>
</tr>
</tbody>
</table>
Various emotions in the two categories can be classified at five levels according to its strengths of subjective feelings as shown in Table 2 (Wang, 2005), where each level encompasses a pair of positive/negative or pleasant/unpleasant emotions.

Definition 3: The strength of emotion $|E_m|$ is a normalized measure of how strong a person’s emotion on a five-level scale identified from 0 through 4, that is:

$$0 \leq |E_m| \leq 4 \quad (1)$$

where $|E_m|$ represents the absolute strength of an emotion regardless whether it is positive (pleasant) or negative (unpleasant), and the scope of $|E_m|$ is corresponding to the definitions of Table 2.

It is observed that an organ known as hypothalamus in the brain is supposed to interpret the properties or types of emotions in terms of pleasant or unpleasant (Payne & Wenger, 1998; Pinel, 1997; Smith, 1993; Wang et al., 2006; Westen, 1999).

Definition 4: Let $T_e$ be a type of emotion, $ES$ the external stimulus, $IS$ the internal perceptual status, and $BL$ the Boolean values true or false. The perceptual mechanism of the hypothalamus can be described as a function, that is:

$$T_e : ES \times IS \rightarrow BL \quad (2)$$

It is interesting that the same event or stimulus $ES$ may be explained in different types, in terms of pleasant or unpleasant, due to the difference of the real-time context of the perceptual status $IS$ of the brain. For instance, walking from home to the office may be interpreted as a pleasant activity for one who likes physical exercise, but the same walk due to car breakdown will be interpreted as unpleasant. This observation and the taxonomy provided in Tables 1 and 2 leads to the following Theorem.

Theorem 1: The human emotional system is a binary system that interprets or perceives an external stimulus and/or internal status as pleasant or unpleasant.

Although there are various emotional categories in different levels, the binary emotional system of the brain provides a set of pairwise universal solutions to express human feelings. For example, angry may be explained as a default solution or generic reaction for an emotional event when there is no better solution available; otherwise, delight will be the default emotional reaction.

<table>
<thead>
<tr>
<th>Level (Positive/Negative)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 No emotion</td>
<td>-</td>
</tr>
<tr>
<td>1 Weak emotion</td>
<td>Comfort, safeness, contentment, fulfillment, trust</td>
</tr>
<tr>
<td></td>
<td>Fear, worry, horror, jealousy, frightening, threatening</td>
</tr>
<tr>
<td>2 Moderate emotion</td>
<td>Joy, delight, fun, interest, pride</td>
</tr>
<tr>
<td></td>
<td>Sadness, anxiety, loneliness, regret, guilt, grief, sorrow, agony</td>
</tr>
<tr>
<td>3 Strong emotion</td>
<td>Pleasure, happiness, bliss, excitement, ecstasy</td>
</tr>
<tr>
<td></td>
<td>Anger, annoyance, hostility, contempt, infuriated, enraged</td>
</tr>
<tr>
<td>4 Strongest emotion</td>
<td>Love, intimacy, passion, amorousness, fondness, infatuation</td>
</tr>
<tr>
<td></td>
<td>Hate, disgust, detestation, abhorrence, bitterness</td>
</tr>
</tbody>
</table>
THE MATHEMATICAL MODEL OF MOTIVATIONS

Motivation is an innate potential power of human beings that energizes behavior. It is motivation that triggers the transformation from thought (information) into action (energy). In other words, human behaviors are the embodiment of motivations. Therefore, any cognitive behavior is driven by an individual motivation.

Definition 5: A motivation is a willingness or desire triggered by an emotion or external stimulus to pursue a goal or a reason for triggering an action.

As described in LRMB (Wang et al., 2006), motivation is a cognitive process of the brain at the perception layer that explains the initiation, persistence, and intensity of personal emotions and desires, which are the faculty of conscious, deliberate, and voluntary choices of actions.

Motivation is a psychological and social modulating and coordinating influence on the direction, vigor, and composition of behavior. This influence arises from a wide variety of internal, environmental, and social sources, and is manifested at many levels of behavioral and neural organizations.

The taxonomy of motives can be classified into two categories known as learned and unlearned (Wittig, 2001). The latter is the primary motives such as the survival motives (hunger, thirst, breath, shelter, sleep, and eliminating pain). The former are the secondary motives such as the need for achievement, friendship, affiliation, dominance of power, and relief anxiety, which are acquired and extended based on the primary motives.

Definition 6: The strength of motivation $M$ is a normalized measure of how strong a person’s motivation on a scale of 0 through 100, that is:

$$0 \leq M \leq 100$$ (3)

where $M = 100$ is the strongest motivation and $M = 0$ is the weakest motivation.

It is observed that the strength of a motivation is determined by multiple factors (Westen, 1999; Wilson & Keil, 1999) such as:

a. The absolute motivation $|E_m|$: The strength of the emotion.

b. The relative motivation $E - S$: A relative difference or inequity between the expectancy of a person $E$ for an object or an action towards a certain goal and the current status $S$ of the person.

c. The cost to fulfill the motivation $C$: A subjective assessment of the effort needed to accomplish the expected goal.

Therefore, the strength of a motivation can be quantitatively analyzed and estimated by the subjective and objective motivations and their cost as described in the following theorem.

**Theorem 2:** The strength of a motivation $M$ is proportional to both the strength of emotion $|E_m|$ and the difference between the expectancy of desire $E$ and the current status $S$, of a person, and is inversely proportional to the cost to accomplish the expected motivation $C$, that is:

$$M = \frac{2.5 \cdot |E_m| \cdot (E-S)}{C}$$ (4)

where $0 \leq |E_m| \leq 4$, $0 \leq (E,S) \leq 10$, $1 \leq C \leq 10$, and the coefficient 2.5 makes the value of $M$ normalized in the scope of [0 .. 100].

In Theorem 2, the strength of a motivation is measured in the scope $0 \leq M \leq 100$. When $M > 1$, the motivation is considered being a desired motivation, because it indicates both an existing emotion and a positive expectancy. The higher the value of $M$, the stronger the motivation.

According to Theorem 2, in a software engineering context, the rational action of a manager of a group is to encourage individual emotional desire, and the expectancy of each software en-
engineer and to decrease the required effort for the employees by providing additional resources or adopting certain tools.

**Corollary 1:** There are super strong motivations toward a resolute goal by a determined expectancy of a person at any cost.

It is noteworthy that a motivation is only a potential mental power of human beings, and a strong motivation will not necessarily result in a behavior or action. The condition for transforming a motivation into a real behavior or action is dependent on multiple factors, such as values, social norms, expected difficulties, availability of resources, and the existence of alternative goals.

The motivation of a person is constrained by the attitude and decision-making strategies of the person. The former is the internal (subjective) judgment of the feasibility of the motivation, and the latter is the external (social) judgment of the feasibility of the motivation. Attitude and decision-making mechanisms will be analyzed in the following subsections.

### THE MATHEMATICAL MODEL OF ATTITUDES

As described in the previous section, motivation is the potential power that may trigger an observable behavior or action. Before the behavior is performed, it is judged by an internal regulation system known as the attitude.

Psychologists perceive attitude in various ways. Fazio (1986) describes an attitude as an association between an act or object and an evaluation. Eagly and Chaiken (1992) define that an attitude is a tendency of a human to evaluate a person, concept, or group positively or negatively in a given context. More recently, Wittig (2001) describes attitude as a learned evaluative reaction to people, objects, events, and other stimuli. Attitudes may be formally defined as follows.

**Definition 7:** An attitude is a subjective tendency towards a motivation, an object, a goal, or an action based on an intuitive evaluation of its feasibility.

The modes of attitudes can be positive or negative, which can be quantitatively analyzed using the following model.

**Definition 8.** The mode of an attitude $A$ is determined by both an objective judgment of its conformance to the social norm $N$ and a subjective judgment of its empirical feasibility $F$, that is:

$$A = \begin{cases} 1, & N = 1 \land F = 1 \\ 0, & N = 1 \lor F = 1 \end{cases} \quad (5)$$

where $A = 1$ indicates a positive attitude; otherwise, it indicates a negative attitude.

### INTERACTIONS BETWEEN MOTIVATION AND ATTITUDE

This section discusses the relationship between the set of interlinked perceptual psychological processes such as emotions, motivations, attitudes, decisions, and behaviors as formally modeled in the preceding sections. A motivation/attitude-driven behavioral model will be developed for formally describing the cognitive processes of motivations and attitudes.

It is observed that motivation and attitude have considerable impact on behavior and influence the ways a person thinks and feels (Westen, 1999). A reasoned action model is proposed by Fishbein and Ajzen (1975) that suggests human behavior is directly generated by behavioral intensions, which are controlled by the attitude and social norms. An initial motivation before the judgment by an attitude is only a temporal idea; with the
judgment of the attitude, it becomes a rational motivation (Wang et al., 2006), also known as the behavioral intention.

The relationship between an emotion, motivation, attitude, and behavior can be formally and quantitatively described by the Motivation/Attitude-Driven Behavior (MADB) model as illustrated in Figure 1. In the MADB model, motivation and attitude have been defined in Equations 4 and 5. The rational motivation, decision, and behavior can be quantitatively analyzed according to the following definitions. It is noteworthy that, as shown in Figure 1, a motivation is triggered by an emotion or desire.

Definition 9: A rational motivation $M_r$ is a motivation regulated by an attitude $A$ with a positive or negative judgment, that is:

$$M_r = M \cdot A = \frac{2.5 \cdot |\varepsilon_m| \cdot (E-S)}{C} \cdot A$$  \hspace{1cm} (6)

Definition 10: A decision $D$ for confirming an attitude for executing a motivated behavior is a binary choice on the basis of the availability of time $T$, resources $R$, and energy $P$, that is:

$$D = \begin{cases} 1, & T \land R \land P = T \land R \land P = 1 \\ 0, & T \lor R \lor P = T \lor R \lor P = 0 \end{cases}$$  \hspace{1cm} (7)

Definition 11. A behavior $B$ driven by a motivation $M_r$ and an attitude is a realized action initiated by a motivation $M$ and supported by a positive attitude $A$ and a positive decision $D$ toward the action, that is:

$$B = \begin{cases} 1, & M_r \cdot D = \frac{2.5 \cdot |\varepsilon_m| \cdot (E-S)}{C} \cdot A \cdot D > 1 \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (8)

**FORMAL DESCRIPTION OF COGNITIVE PROCESSES OF MOTIVATION AND ATTITUDE**

The formal models of emotion, motivation, and attitude have been developed in previous sections. This section extends the models and their relationship into detailed cognitive processes based on the OAR model (Wang, 2007d) and using RTPA (Wang, 2002b, 2003c, 2006b, 2007a), which enable more rigorous treatment and computer simulations of the MADB model.
The Cognitive Process of Motivations

The mathematical model of rational motivation is described in Equation 6. Based on Equation 6, the cognitive process of motivation is presented in Figure 2. The motivation process is divided into four major sub-processes known as: (1) to form motivation goal; (2) to estimate strength of motivation; (3) to form rational motivation; and (4) to stimulate behavior for the motivation.

The MADB model provides a formal explanation of the mechanism and relationship between motivation, attitude, and behavior. The model can be used to describe how the motivation process drives human behaviors and actions, and how the attitude as well as the decision-making process help to regulate the motivation and determines whether the motivation should be implemented.

The Cognitive Process of Attitudes

The mathematical model of attitude has been described in Equation 5. Based on Equation 5, the cognitive process of attitude is presented in Figure 3. The attitude process is divided into three major sub-processes known as: (1) to check the mode of attitude; (2) to determine physical availability; and (3) to stimulate behavior for the motivation.

The Integrated Process of Motivation and Attitudes

According to the MADB model and the formal description of the motivation and attitude processes as shown in Figures 1 through 3, the cognitive processes of motivation and attitude are interleaved. An integrated process that combines both motivation and attitude is given in Figure 4 via the following sub-processes: (1) to form motivation goals; (2) to estimate strength of motivation; (3) to check the mode of attitude; (4) to form rational motivation; (5) to determine physical availability; and (6) to stimulate behavior for the rational motivation.

MAXIMIZING STRENGTHS OF INDIVIDUAL MOTIVATIONS

Studies in sociopsychology provide a rich theoretical basis for perceiving new insights into the organization of software engineering. It is noteworthy that in a software organization, according to Theorem 2, the strength of a motivation of individuals $M$ is proportional to both the strength of emotion and the difference between the expectancy and the current status of a person. At the same time, it is inversely proportional to the cost to accomplish the expected motivation $C$.

The job of management at different levels of an organization tree is to encourage and improve $E_m$ and $E$, and to help employees to reduce $C$.

**Example 1:** In software engineering project organization, the manager and programmers may be motivated to the improvement of software quality to a different extent. Assume the following factors as shown in Table 3 are collected from a project on the strengths of motivations to improve the quality of a software system, analyze how the factors influence the strengths of motivations of the manager and the programmers.

According to Theorem 2, the strengths of motivations of the manager $M_1$ and the programmers $M_2$ can be estimated using Equation 4, respectively:

$$M_{1}(\text{manager}) = \frac{2.5 \cdot | E_m | \cdot (E-S)}{C}$$

$C = 2.5 \cdot 4 \cdot (8 - 5)$

$= 10.0$

and

$$M_{2}(\text{programmer}) = \frac{2.5 \cdot 3.6 \cdot (8 - 6)}{8}$$

$= 2.3$
### The Motivation Process

**Motivation** (I:: oS; O:: OAR ST) \( \triangleq \)

\{ I. Form motivation goal(s) \\
   \rightarrow Object Identification (o, A, R ) \\

II. Estimate strength of motivation \( M(o) \) \\
   \rightarrow Quantify \( E_m(o) \) // The strength of emotion \\
   \rightarrow Quantify \( S(o) \) // The current status \\
   \rightarrow Quantify \( E(o) \) // The expectancy of desire \\
   \rightarrow Quantify \( C(o) \) // The cost to accomplish \\
   \rightarrow \begin{align*} 
   M(o) &= 2.5 \frac{E_m(o) \circ (E(o) \circ S(o))}{C(o)} \\
   \end{align*} \\
   \rightarrow \begin{cases} 
   \text{Positive motivation} & (M(o) > 1) \\
   \text{Negative motivation} & (M(o) < 1) \\
   \end{cases} \\

III. Check the mode of attitude \( A(o) \) \\
   // Refer to the Attitude process \\

IV. Form rational motivation \( M_r(o) \) \\
   \rightarrow M_r(o) := M(o) \odot A(o) \\
   \rightarrow \begin{cases} 
   \text{Rational motivation} & (M_r(o) > 1) \\
   \text{Irrational motivation} & (M_r(o) < 1) \\
   \end{cases} \\

V. Determine physical availability \( D(o) \) \\
   // Refer to the Attitude process \\

VI. Stimulate behavior for \( M_r(o) \) \\
   \rightarrow ( \text{Implement motivation } o ) \\
   \rightarrow GenerateAction (M_r(o)) \\
   \rightarrow ExecuteAction (M_r(o)) \\
   \rightarrow R := R \cup <o, M_r(o)> \\
   \rightarrow D(o) := 0 \\
   \rightarrow o := o \\
   \rightarrow R := \\
   \rightarrow OAR ST = <O \cup o, A \cup A, R \cup R> \ // \text{Form new OAR model} \\
   \rightarrow Memorization (OAR ST)
\}

---

**Figure 2.** The cognitive process of motivation
On the Cognitive Processes of Human Perception with Emotions, Motivations, and Attitudes

Figure 3. The cognitive process of attitude

\[
\text{The Attitude Process}
\]

\[
\text{Attitude (I:: S, O:: OAR ST)} \triangleq \\
\{ \text{I. Form motivation goal(s)} \\
\quad \rightarrow \text{ObjectIdentification (o, A, R)} \\
\text{II. Estimate strength of motivation M(o)} \\
\text{\qquad // Refer to the Motivation process} \\
\text{III. Check the mode of attitude A(o)} \\
\text{\qquad // Perceptual feasibility} \\
\quad \rightarrow \text{Qualify (N(o)BL)} \quad \text{// The social norm} \\
\quad \rightarrow \text{Qualify (F(o)BL)} \quad \text{// The subjective feasibility} \\
\text{\qquad \rightarrow ( } \text{N(o)BL \land F(o)BL = T} \\
\text{\qquad \quad \rightarrow A(o)N := 1} \\
\text{\qquad \quad \rightarrow A(o)N := 0} \\
\text{\qquad )} \\
\text{IV. Form rational motivation Mr(o)} \\
\text{\qquad // Refer to the Motivation process} \\
\text{V. Determine physical availability D(o)} \\
\text{\qquad // Refer to the Motivation process} \\
\text{\qquad \rightarrow \text{Qualify (T(o)BL)} \quad \text{// The time availability} \\
\text{\qquad \rightarrow \text{Qualify (R(o)BL)} \quad \text{// The resource availability} \\
\text{\qquad \rightarrow \text{Qualify (P(o)BL)} \quad \text{// The energy availability} \\
\text{\qquad \rightarrow ( } \text{T(o)BL \land R(o)BL \land P(o)BL = T} \\
\text{\qquad \quad \rightarrow D(o)N := 1} \quad \text{// Confirmed motivation} \\
\text{\qquad \quad \rightarrow D(o)N := 0} \quad \text{// Infeasible motivation} \\
\text{\qquad )} \\
\text{VI. Stimulate behavior for Mr(o)} \\
\text{\qquad \rightarrow ( } \text{D(o)N = 1} \quad \text{// Implement motivation o} \\
\text{\qquad \quad \rightarrow \text{GenerateAction (M(o))} } \\
\text{\qquad \quad \rightarrow \text{ExecuteAction (M(o))} } \\
\text{\qquad \quad \rightarrow R := R \cup <o, M(o)> } \\
\text{\qquad \quad \rightarrow D(o)N := 0} \\
\text{\qquad \quad \rightarrow o := R := } \\
\text{\qquad \quad \rightarrow \text{OAR ST} = <O \cup o, A \cup A, R \cup R> // Form new OAR model} \\
\text{\qquad \rightarrow \text{Memorization (OAR ST)}} \}
\]

Table 3. Motivation factors of a project

<table>
<thead>
<tr>
<th>Role</th>
<th>E_m</th>
<th>C</th>
<th>E</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>The manager</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Programmers</td>
<td>3.6</td>
<td>8</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>
The Motivation and Attitude Process

Motivation-Attitude (I : s$; O : OAR ST) ≠

I. Form motivation goal(s)
   → ObjectIdentification ($o , A' , R'$)

II. Estimate strength of motivation $M(o)$$\|$ (I : o$; O : OAR$)
   → Quantify ($E_o(o)$$\|$) // The strength of emotion
   → Quantify ($S(o)$$\|$) // The current status
   → Quantify ($E(o)$$\|$) // The expectancy of desire
   → Quantify ($C(o)$$\|$) // The cost to accomplish
   → $M(o) = \frac{2.5 \cdot E_o(o) \cdot S(o) \cdot E(o) \cdot C(o)}{C(o)}$

   \[ \begin{align*}
   ( & \ 
   & \text{M}(o) > 1 ~ \Rightarrow ~ @M(o) = T \quad \text{(Positive motivation)} \\
   & \neg \text{M}(o) = F \quad \text{(Negative motivation)}
   \end{align*} \]

III. Check the mode of attitude $A(o)$$\|$ // perceptual feasibility
    → Qualify ($N(o)$$\|$) // The social norm
    → Qualify ($F(o)$$\|$) // The subjective feasibility
    → ($ \neg \text{N}(o) \land \text{F}(o) = T \Rightarrow A(o) = 1$
    \quad \neg \text{A}(o) = 0$

IV. Form rational motivation $M_r(o)$
    → $M_r(o) = M(o) \cdot A(o)$$\|$ (I : o$; O : OAR$)

    \[ \begin{align*}
    ( & \ 
    & \text{M}(o) > 1 ~ \Rightarrow ~ @M_r = T \quad \text{(Rational motivation)} \\
    & \neg \text{M}_r = F \quad \text{(Irrational motivation)}
    \end{align*} \]

V. Determine physical availability $D(o)$$\|$ // resource availability
    → Qualify ($T(o)$$\|$) // The time availability
    → Qualify ($R(o)$$\|$) // The resource availability
    → Qualify ($P(o)$$\|$) // The energy availability
    → ($ T(o) \land R(o) \land P(o) = T \Rightarrow D(o) = 1$
    \quad \neg D(o) = 0$

VI. Stimulate behavior for $M_r(o)$
    → ($ D(o) = 1 \Rightarrow \text{Implement motivation o}$
    \quad \Rightarrow \text{GenerateAction} (M_r(o))
    \quad \Rightarrow \text{ExecuteAction} (M_r(o))
    \quad \Rightarrow R' := R' \cup <o, M_r(o)>
    \quad \neg D(o) = 0$
    \quad \Rightarrow A' := \emptyset
    \quad \Rightarrow R' := \emptyset$

\[ \text{OAR ST = } O \cup o , A \cup A', R \cup R' \Rightarrow \text{Form new OAR model} \]
\[ \Rightarrow \text{Memorization} (\text{OAR ST}) \]

---

**Figure 4. The integrated process of motivation and attitude**
The results show that the manager has much stronger motivation to improve the quality of the software system than that of the programmers in the given project. Therefore, the rational action for the manager is to encourage the expectancy of the programmers or to decrease the required effort for the programmers by providing additional resources or adopting certain tools.

According to sociopsychology (Wiggins, Eiggins, & Zanden, 1994), social environment, such as culture, ethical norms, and attitude greatly influences people's motivation, behavior, productivity, and quality towards collaborative work. The chain of individual motivation in a software organization can be illustrated as shown in Figure 5.

Cultures and values of a software development organization helps to establish a set of ethical principles or standards shared by individuals of the organization for judging and normalizing social behaviors. The identification of a larger set of values and organizational policy towards social relations may be helpful to normalize individual and collective behaviors in a software development organization that produces information products for a global market.

Another condition for supporting creative work of individuals in a software development organization is to encourage diversity in both ways of thinking and work allocation. It is observed in social ecology that a great diversity of species and a complex and intricate pattern of interactions among the populations of a community may confer greater stability on an ecosystem.

**Definition 12:** Diversity refers to the social and technical differences of people in working organizations.

Diversity includes a wide range of differences between people such as those of race, ethnicity, age, gender, disability, skills, educations, experience, values, native language, and culture.

**Theorem 3:** The diversity principle states that the more diversity of the workforce in an organization, the higher the opportunity to form new relations and connections that leads to the gain of the system fusion effect.

Theorem 3 is particularly useful for software development organizations where creative work products are engineered. System theory indicates that if the number of components of a system reaches a certain level—the critical mass—then the functionality of the system may be dramatically increased (Wang, 2007a). That is, the increase of diversity in a system is the condition to realize the system fusion effect, which results in a more powerful system with newly created relations and behaviors that only belong to the system as a whole.
CONCLUSION

This article has described the perceptual processes of emotions, motivations, and attitudes based on which complicated psychological and mental processes may be formally modeled and rigorously explained. Relationships and interactions between motivation and attitude have been formally described in RTPA. It has been recognized that the human emotional system is a binary system that interprets or perceives an external stimulus and/or internal status as in the categories of pleasant or unpleasant. It has revealed that the strength of a motivation is proportional to both the strength of the emotion and the difference between the expectancy of desire and the current status of a person and is inversely proportional to the cost to accomplish the expected motivation. Case studies on applications of the interactive motivation-attitude theory and cognitive processes of motivations and attitudes in software engineering have been presented.

This work has demonstrated that the complicated human emotional and perceptual phenomena, as well as their natural drives and constraints, can be rigorously modeled in denotational mathematics and be formally treated and described. This work has been based on two fundamental cognitive informatics models: the LRMB and the OAR model. The former has provided a blueprint to exploring the natural intelligence and its mechanisms. The latter has established a contextual foundation to reveal the logical representation of information, knowledge, and skills in the perceptual space of the brain.

ACKNOWLEDGMENT

The author would like to acknowledge the Natural Science and Engineering Council of Canada (NSERC) for its support to this work. We would like to thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES


Wang, Y. (2003c). Using process algebra to describe human and software behaviors. Brain...


Chapter 2.21
Sociotechnical System Design for Learning: Bridging the Digital Divide with CompILE

Benjamin E. Erlandson
Arizona State University, USA

ABSTRACT

CompILE is a sociotechnical “comprehensive interactive learning environment” system for personal knowledge management and visualization that represents the growing collective knowledge an individual gathers throughout his or her lifespan. A network of intelligent agents connects the user and his or her inhabited knowledge space to external information sources and a multitude of fellow users. Following a brief perspective on educational technology, concepts of human-computer interaction, and a description of CompILE, this chapter will introduce CompILE as a sociotechnical system supported by an enriched design process. From an educational perspective, CompILE can bridge the digital divide by creating community, embracing culture, and promoting a learning society.

INTRODUCTION

This chapter begins with a brief perspective on educational technology, concepts of human-computer interaction, and a description of the Comprehensive Interactive Learning Environment (CompILE) as a knowledge management system controlled using a network of intelligent software agents. Strategies for bridging the digital divide using CompILE will be presented using the four-part framework of the “Wheel of Policy Instruments” constructed by van Dijk (2005).

The effect of CompILE on society will be approached from the standpoint of lifelong learning and usability engineering. CompILE will create a sense of community via participatory research, scenario-based design, and heuristic evaluation. In terms of culture, CompILE will rely heavily
upon ethnography to refine its ability to facilitate cross-cultural communication, and a distinct effort to achieve functional beauty as an interactive system should help its participants achieve a higher level of appreciation for the aesthetics of “culture” from a philosophical perspective. Globalization can be addressed by CompILE in terms of ubiquitous computing, the mobile user, visual literacy, and just-in-time translation. The chapter will conclude with a description of CompILE as a hybrid entity, and the essence of CompILE as a sociotechnical system.

This chapter has two specific objectives:

• Introducing CompILE as a sociotechnical system supported by an enriched design process
• Using CompILE to bridge the digital divide by creating community, embracing culture, and promoting a learning society.

BACKGROUND

One trait of humans that makes us advanced social creatures is our ability to create artificial devices or artifacts that expand our capabilities (Norman, 1993). Throughout history, many of these artifacts have been perceived as the final piece to the ultimate puzzle of technological advancement; the last thing humans would create to meet all their future needs. The automobile, the airplane, the telephone, the television, the microwave, the Internet; the list goes on. While no single invention can actually be that final puzzle piece, major technological advancements cannot be denied as a driving force of society. Norman (1993) agrees, noting the essential nature of technology for growth of human knowledge and mental capabilities.

An important consideration to be made when designing any new interface is the anticipated impact it will have on human activity. In fact, most significant consequences of design are based on impact to human activity. In an effort to reduce occurrence of these consequences, the design process should include a well-managed team of participants with a variety of knowledge and skills, willing to completely deconstruct any particular problem that may arise (Carroll, 1991). Why are these consequences so important? These consequences have the power to make technology completely redundant, existing merely for entertainment.

This breakpoint of redundancy is paramount in education, which, upon clarification, is essentially a purposed form of communication between one or more individuals, intended to facilitate learning and knowledge construction. Using a learning-centric approach, the application of technology to the processes of teaching and learning can be seen as an amplification of the channels of communication. There is no point in seeking to improve the educational communication process with technology if all efforts result in glamorous toys with no effect beyond distraction. Computers cannot be placed in schools for an administration to show that it is on the forefront with cutting-edge educational technology. Technology must be carefully integrated with curricula to be used as tools, or artifacts, fundamental to the learning process.

This concept extends beyond the classroom geographically and temporally. Students using the technology must be wholly engaged in the content of the curriculum, using the technology as a mechanism of that engagement. Kling (2000) reinforces this notion of engagement when describing such an application of technology as the development of a “sociotechnical system,” or an information technology system that cannot separate its existence from the social context of its users. Once this point of engagement is reached, the technology available today and tomorrow can create amazing opportunities for experience and growth within and beyond the classroom. To reach this point, a sociotechnical system must be designed and developed, with an eye toward
the inseparable bond between human, machine, and use.

**HUMAN-COMPUTER INTERACTION**

There are four main sources of knowledge underlying information systems design: (1) systematic empirical evidence of user-interface relations, (2) models and theories of human cognition and artificial systems, (3) field studies of how humans cope with tasks in real-world settings, and (4) specific work domain knowledge derived from textbooks, interviews with experts, and analytical methods such as task analysis (Rasmussen, Anderson, & Bernsen, 1991). The diversity of these sources underscores the interdisciplinary nature of the studies and practices involved with human-computer interaction (HCI), including but not limited to the natural sciences, engineering, the social sciences, and humanities.

One might assume that the foundations of HCI are purely objective, due to the binary nature of information systems. Computers are objective machines based on logic, but humans are interacting with them; this completely unbalances the equation. HCI research and development must take into account human knowledge, decision-making and planning, current mental representations of the user based on causal and intentional relationships, and the fact that interactive behavior is influenced by the environment in which it takes place (Beehtel, 1997; Gifford, 1997; Sommer, 1969).

It is increasingly inadequate for system designers to consider the modules of the information system in isolation from each other (Rasmussen et al, 1991). Increasingly numerous applications working in seamless integration on the same computer (and across networks) becomes more feasible with each new cycle in computer processing technology. Modules can no longer be designed around the requirements of a narrow group of users (Rasmussen et al, 1991). As greater numbers of users become cognizant of the true potential of personal computing, the “narrow” user groups formerly associated with modules and applications in information systems are expanding rapidly. The expansion presents an enormous challenge to designers; accounting for so much more variety when building interfaces for their products. Accounting for variety involves the careful specification of user communities in tandem with a thorough identification of tasks. With novice users, the most success with this dual process can be found through reliance on several measurable criteria: time taken to learn specific functions, speed of task performance, rate of human error, subjective user satisfaction, and human retention of functions over time (Shneiderman, 1993).

The less time users must spend learning a system, the more time they can spend being productive. In man-made environments (interfaces), there are two phases of activity: adaptation and adapted. During adaptation, a user’s activity is being guided by exploratory behavior. Once adapted, a user’s behavioral patterns have reached a state of equilibrium, or a *satisfactory state of transparency* between user input and system output (Rasmussen et al., 1991). The characteristics of any particular workplace and the type of work being done determine the priority of the two phases.

**COMPREHENSIVE INTERACTIVE LEARNING ENVIRONMENTS**

The “workplace” of most interest to those designing interfaces for educational technology would be the transitioning traditional classroom, which can be propelled beyond the 21st century with CompIILE. CompIILE relies heavily upon provision of access to information to all students at all stages of the educational process, as well as continual facilitation of mediated communication between humans and machines. This is accom-
plished through a network of intelligent agents whose existence is focused entirely upon two processes: (1) facilitation of learning for the user and (2) constantly learning about the user. These two processes are inseparable, as an integral part of the agents' facilitation of learning processes for each user is learning the “quirks” of each user’s processes in situations of learning (both formal and informal). To create a personal knowledge management system, agents must know the personality of the user almost better than the user knows himself or herself.

As a system for personal knowledge management and visualization, CompILE can be considered an inhabited information space, or IIS (Snowdon, Churchill, & Frécon, 2004), that represents the growing collective knowledge an individual gathers throughout his or her lifespan. To be more specific, CompILE connects the user and his or her inhabited knowledge space (IKS). This IKS can be navigated individually by the user or collaboratively with additional individuals. Any number of interfaces could be used to access this IKS. Most of these interfaces would function as intermediaries between the internal knowledge architecture, external information sources, the user(s), and associated intelligent agents.

These agents can permanently link education and technology into a symbiotic relationship. Agents have become fairly reliable forms of autonomous software, intertwined more actively with the social lives of humans than any technology ever before, playing “an increasingly critical role in the organization, in particular the self-organization, of social life” (Brown & Duguid, 2000). Education is a highly socialized activity, with constant interaction between people a fundamental part of learning. A positive attribute of software agents is their ability to deal with the increasingly intimidating amounts of information that people are forced to deal with on a daily basis (Brown & Duguid, 2000). This ability would enable agents to assist students with keeping track of much larger amounts of information.

Agents that exist today (mainly for consumer purposes) can be broken down into three categories: information brokering, product brokering, and merchant brokering. Information brokers are agents that gather information from a wide variety of resources, and perform operations with the gathered information. Product broker agents are designed to track their owner’s history (usually shopping habits) and alert that person to the availability of particular products in various venues. Agents that perform merchant brokering are essentially price comparison agents, trying to find the best possible “buy” for their owner. No successful negotiation agents (a fourth category) have ever been created (Brown & Duguid, 2000). Humans negotiate through nonverbal behavior in order to accommodate themselves to one another; constructing and adjusting the social fabric, but software agents have yet to be built with this capacity.

Information brokering agents would have the most direct crossover into education, being resourceful for students, teachers, and administrators; taking the form of research agents. Product brokering agents would translate to learning agents, tracking learning styles and suggesting individual supplements and alternative approaches to the basic curriculum for each student. Merchant brokering agents would not necessarily be feasible in an educational environment on a regular basis, but a similar concept can be realized in terms of brokering a user’s time. Time management agents can assist the user by calculating approximate durations for activities, with a focus on multitasking and efficiency. Negotiation agents are an integral part of the CompILE system, facilitating intercommunication amongst all types of participants, with the agents dutifully negotiating with each other on the users’ behaviors as the users themselves go about their daily routines. As a networked system of these four types of agents, CompILE could prove to be a vital instrument to help our society bridge the “digital divide.”
THE DIGITAL DIVIDE: WHEEL OF POLICY INSTRUMENTS

Van Dijk (2005) addresses the “digital divide” with a four-part strategy called the “Wheel of Policy Instruments,” which denotes four perspectives for access to information and communication technology (ICT): motivational, material, skills, and usage. Motivational access features increasing surplus values and user-friendliness of ICTs, organizing user trust by regulation, promoting useful applications of ICTs, and providing specific services for underserved groups. In terms of surplus value and user-friendliness, every use of CompILE must be beneficial for the user, with little or no margin for “user errors” that do nothing but frustrate the user, decreasing user confidence in the reliability of the CompILE system. Additionally, certain agents can focus upon the “add-on value” of a user’s current actions within CompILE, giving suggestions for efficiency that the user may or may not have noticed; thus, increasing the surplus value of the system.

Regulation and promotion of ICTs can be seen as a symbiotic relationship. Government and corporate institutions can work together to promote useful applications of CompILE by upholding regulations in an open, trustworthy manner that promotes the benefits of CompILE (and other applications and services) to the user. A democratic, user-centric approach to legislation of regulations can serve as a valuable promotional tool as well. CompILE can help eliminate “underserved groups” using special agents designed to find information and services specifically targeted to the individual demographics of the agents’ owner/user, such as cultural groups, gender- and age-specific information, or sociopolitical, health, and spiritual or secular information.

Material access features universal service, access to basic provisions, promotion of broadband access, support of competition, interconnectivity amongst media, conditional subsidies to lagging groups, and creation of public access points. As insinuated, much of materials access improvement involves government and corporate institutions cooperating to provide equal access to ICT services for all citizens. A good start would be the deprivatization of broadband Internet access into a public utility, much like public water and sewer service or electricity co-ops. Furthermore, public access points for broadband Internet connectivity (both wired and wireless) would be a fundamental element for the complete operation of CompILE. A combination of passive RFID tags and agents would allow for user identity and seamless integration from one access point to another, with just-in-time delivery of information to the user. This should all be available free of charge as part of a public broadband infrastructure.

Skills access features connecting all schools, libraries, hospitals, and cultural centers, adapting curricula for operational, instrumental, and strategic skills, creating better educational software and better digital teaching skills, teaching both basic and advanced skills to everyone, extending adult education to learning digital skills, and an emphasis upon distance education at home and in local classes or meetings. Adaptation of curricula for operational, instrumental, and strategic ICT skills would help users “meet” advancing technology at a “middle ground” of usability and user-friendliness. This should also lead to an increased knowledge base and skill sets for those who will be maintaining and supporting the CompILE system on a regular basis.

CompILE can reduce the need for purposed standalone educational software applications by providing such a dynamic, comprehensive digital teaching and learning environment. Additionally, CompILE can improve digital teaching skills with the assistance of special agents for preservice teachers while they are still in training. These agents can be carried over with these teachers into the in-service environment for professional development to expand technological skills and knowledge.
As opposed to a somewhat narrowed emphasis on “distance education” at home and in local classes, an emphasis should be made on a broader inclusion of self-paced, asynchronous collateral/supplemental learning and inquiry. The path of learning should extend throughout the waking hours of each individual’s day. The formation of networked learning groups could supplement the current formal education system, leading to a “cottage industry” of sorts in the field of education and training. CompILE could facilitate these encounters both synchronously and asynchronously, both in-person and across distances via ICTs.

Usage access features support for learning on the job and at home, more hardware and software designed for underserved groups, special content for minorities and socially deprived groups, open access to all public and scientific information and major cultural events, support for competition in software and services, promotion of open-source software, lifelong learning and circulation of applications in jobs and education, and the full integration of ICTs into social and user environments. The success of CompILE depends (perhaps most heavily) upon open access to all public and scientific information. This information must be available to all who wish to learn, reflecting back upon the users’ trust of the system—CompILE’s agents must be able to deliver all the pertinent information related to a subject of inquiry. Also, upon completion of an initial performance or showing, major cultural events should be released into the public domain. The option should be available for immediate, free access to any cultural event, in appropriate digital format via the Internet.

The CompILE application will be open source software. With the source code available for public development, advanced users would be able to modify CompILE with add-ons or “plug-ins” that can increase the benefits of CompILE use. Similar to other open source software projects, these add-ons will be shared with other users across the Internet. In addition to extending the basic internal functionality of the main CompILE application, the add-ons would allow CompILE to communicate with an increasing number of current (and future) software applications. This would allow for a much fuller integration of ICTs into social and user environments, based on the intended widespread use of CompILE as a tool set for learning. The increased user-friendliness and surplus value realized with better communication between CompILE, and other applications would promote lifelong learning and the circulation of useful software applications in jobs and education.

CompILE AND SOCIETY

The concept of “lifelong learning” has its official origination in the 1973 UNESCO “Faure Report” (Faure, Herrera, Kaddoura, Lopos, Petrovsky, Rahnema, & Ward, 1972), which indicates the world’s transition to a “learning society” and a “knowledge economy.” Since then, many definitions of lifelong learning have surfaced. Smith and Spurling (1999) approach lifelong learning from a holistic perspective, with empirical and moral elements. Concerned with the scope of lifelong learning, the empirical element includes continuity, intention, and unfolding strategy as keys to the learning process. The moral element deals with the character of the learner, in terms of personal and social commitment, respect of others, and respect of the truth. Most importantly, Smith and Spurling give a wide berth to the activities that fall within the realm of lifelong learning, including formal, informal, and self-directed learning.

Crick (2005) hints toward the autonomy of the learner, noting that learning and metacognition are activities that we must do for ourselves, leading to a stronger self-awareness of and responsibility for learning. Fischer (2001) maintains that the arena for lifelong learning is highly contextualized and driven by the specific yet flexible demands of
these situations of context. Considering the wide range of learning activities, learner autonomy, and highly contextual situations of learning, it’s no surprise that humans must embrace the concept of a “learning society.”

One approach to understanding this concept of a “learning society” is in terms of a social learning system, which can be understood as a “constellation” of communities of practice (Wenger, 2000). However, from a sociopolitical perspective, Lambeir (2005) suggests that “lifelong learning has become a new kind of power mechanism,” highlighting two important trends: (1) a shift from a “knowledge-based” society to an information society, with learning having an emphasis upon the retrieval, dissemination, and evaluation of information; and (2) the increasing autonomy of the learner. With or without consideration of the politics involved, how can technology be applied to such a socially expansive concept?

The Open University Knowledge Network is under development as a software environment for lifelong learners’ community knowledge network, based on Wenger’s community of practice framework (McAndrew, Clow, Taylor, & Aczel, 2004). The designers chose to use evolutionary design methods, building only essential elements into the software, to be improved with user feedback. From their experiences with the OU Knowledge Network, they conclude that knowledge management software is an excellent tool for the collection of resources and experiences within a community, only if great care is taken when introducing the system and in maintaining existing work practices.

Koper and Tattersall (2004) note four main characteristics of lifelong learning to be considered when developing ICT support networks:

- Self-direction of learners
- Mobile access to distributed information
- Heterogeneity of learners
- Necessity for the maintenance of individual records of growth in competence

Their “Learning Network” (Koper & Sloep, 2003), developed exclusively with open source software, is currently in the pilot stages, also at the Open University of the Netherlands. Designed to support “seamless, ubiquitous access” to learning facilities at work, home, and schools, this architecture helps lower the access barriers to lifelong learning (Koper & Tattersall, 2004).

Two crucial areas of research taking place simultaneously at the Center for Lifelong Learning and Design (L3D) are knowledge-based systems (KBS) and human-computer communication (HCC), with the idea that success cannot fully be realized in either area without substantial progress in the other (Fischer, 2001). All of these projects surrounding the application of technology in support of lifelong learning have similar indications, working with information, autonomy and heterogeneity of the learner, improvement through user feedback, maintaining existing work practices, and the inseparability of KBS and HCC, lead to a necessary focus upon the absolute importance of the usability of applied technology.

**USABILITY**

There are five distinct attributes of usability: learnability, efficiency, memorability, errors, and satisfaction. An interface should be simple enough to learn that users can quickly become productive with the application. Efficiency is also essential, as a user should be able to maintain high levels of productivity with the application upon learning the interface. The interface should be easy to remember, so those users that are not constantly in contact with the application can easily reorient themselves when necessary. User error should be kept to a minimum, and errors that are unavoidable should have easy avenues of recovery. The interface should be enjoyable for the user, creating a sense of satisfaction with the application (Nielsen, 1993).
The “common sense” foundation of usability makes it quite easy for a designer to forget the actual level of importance that usability has in the design process. The concepts become second nature, and because of this may become “second-rate.” Also, it may seem as though an interface designer for a fairly complex system would have to choose between ease of use (learnability) and productivity (efficiency). This choice can be avoided by providing multiple interaction styles along with user action accelerators (Nielsen, 1993). Through a purposed focus upon multiple interaction styles, accelerators, or both, CompILE must compensate completely for learnability and efficiency. Not only should this focus increase endearment of the user population to CompILE, it should create a general increase in the appreciation of the usability design process amongst members of this user community.

**CompILE AND COMMUNITY**

The discovery-based design process inherent in sociotechnical systems effectively integrates the designers as participants within the system itself. This integration indicates a need for change within educational research as well. Geelan (2006) would agree, noting that the time has come for drastic changes in the methods of educational research, with the need for a highly participatory approach. This can be accomplished with an appropriate combination of scenario-based design, heuristic evaluation, and a better understanding of individual human participatory roles within CompILE.

**SCENARIO-BASED DESIGN**

While most engineering methods seek to control the complexity and fluidity of design using filtering and decomposition techniques, scenario-based design techniques seek to exploit the complexity and fluidity of design by learning more about the structure and dynamics of the problem domain. This process of exploitation promotes diverse views of the situation at hand, as well as intimate interaction with the concrete elements of that problem domain (Carroll, 2000). To exemplify this scenario-based process, consider an ambitious educational administration that wishes to implement CompILE across its entire school system. The problem domain of CompILE consists of all the individual schools within the system and the “usual” issues that these separate schools face: curricular consistency, communication, funding, staffing, regulations, parental involvement, and so forth.

There are two fundamental premises of the scenario-based approach that make it such an attractive design option. First, descriptions of how people accomplish tasks are a primary working design representation. Second, each context in which humans experience and act provides detailed constraint for the development and application of information technologies (Carroll, 2000). Considering these two premises, it is easy to see how a scenario-based approach for designing CompILE would radically differ from the same type of approach for designing an accident prevention program for a nuclear power plant.

Scenarios are contexts of interaction that highlight key aspects of a particular problem domain or system: goals suggested by the appearance and behavior of the system; what people try to do with the system; what procedures are adopted or not adopted, carried out successfully or erroneously; and reflections of system participants. In addition, these scenarios are comprised of four characteristic elements: setting, or the place and its props; agents/actors, or the people operating within the system; goals/objectives, or the changes an agent/actor wishes to achieve in the circumstances of the setting; and plot, including sequences of actions and events that can facilitate, obstruct, or be irrelevant to the given goals/objectives (Carroll, 2000).
Creating and using scenarios pushes designers beyond simple static answers. The emphasis on raising questions makes it easier for designers to integrate reflection and action into their own design practice. The process creates constant integration between the designers and the constituents of the problem domain by evoking reflection, contrasting the simultaneously concrete and flexible nature of scenarios, promoting work orientation, and melding abstraction with categorization. Scenarios also allow analysts and designers to visually sketch interactions in order to probe relationships (Carroll, 2000). It is precisely the nuances revealed through these questions and sketches that hold the keys to progress within the design of a system like CompILE. In addition, much of the reflection/action paradox is resolved by scenarios, as they provide a language for action that invokes reflection (Carroll, 2000). For example, a designer concerned with the implementation of stand-alone Web-browsing functionality within the global navigation menu of the CompILE interface would have a much easier time making a primary decision if he or she knew the current classroom dynamic between students, teachers, and the integration of Web content as a daily part of the curriculum.

Scenarios of use reconcile concreteness and flexibility by embodying concrete design actions that evoke concrete move testing by designers, and by facilitating open-ended exploration of design requirements and possibilities (Carroll, 2000). This approach is perfect for a project such as CompILE, because many decisions would be cut-and-dry, based on preceding projects within the same vein, but the specific needs of any school system would call for the refreshment of open-ended exploration. Exploration of this nature would help avoid overapplication of technology, as would the practical approach to interface use and analysis; heuristic evaluation.

**HEURISTIC EVALUATION**

Heuristic evaluation is the process of a person viewing an interface and making value judgments based on the 10 heuristics of usability, using his or her own common sense or intuition. Nielsen (1993) provides the following 10 heuristics of usability:

- Simple and natural dialog
- Speak the users’ language
- Minimize user memory load
- Consistency
- Feedback
- Clearly marked exits
- Shortcuts
- Good error messages
- Error prevention
- Help and documentation

One of the best ways to find mistakes and problems in interfaces is to “use” the interface and look for them. Evaluators should work individually, only communicating after completion, including written or recorded reports following inspection (Nielsen, 1993). This lowers the probability of biased evaluations. For CompILE, individualized evaluation might not work, since students would be using the interface in close proximity to each other within the classroom environment. Even though the results would likely be biased for these group evaluation sessions, the bias itself would be an integral part of the feedback loop in designing CompILE. Examining “off-line” communication between students would be essential for creating seamless computerized intercommunication functionality.

Observers are a necessity for evaluating CompILE, since the evaluators would include young children, incapable of successfully completing comprehensive written evaluations. However, the “observer method” would still have its shortcomings, as much of the responsibility for interface content analysis still falls upon the evaluator. Due
to the inexperience that most children would have with such an evaluation, the fact that observers are free to give ample hints for interface operation would be crucial to successful testing of CompILE. The short attention spans of most school-aged children combined with the complex nature of the CompILE interface would necessitate the series of short, small-scope evaluation sessions (as opposed to one marathon session). The product of a heuristic evaluation of any interface is a list of problems about that interface, directly referencing the 10 heuristics. When designers receive the report, the dots between problem and solution are mostly connected.

Unfortunately, many children might have difficulty remembering the 10 heuristics. There are at least three solutions to compensate for this difficulty: (1) place posters or reference cards of the heuristics near the child evaluator; (2) provide simple training for any child who would be heuristically evaluating the interface; or (3) create a third role within the evaluation process, a “supporter.” A supporter would be an observer with more “clout,” helping the child with his or her analysis as opposed to merely recording it. However, this supporter role does not have to be entirely fulfilled by humans. Supporters could manifest as “quality feedback” agents, similar to many software applications already in existence. Instead of the typical reactionary role that most of these quality feedback agents play, CompILE’s supporter agents would assume a highly proactive role, providing an almost constant link between the users (from all roles of participation) and the processes of scenario-based design and heuristic evaluation.

**ROLES OF PARTICIPATION**

How can these agents specifically benefit the current and future individuals involved in the educational system: students, teachers, parents, administrators, researchers, and designers? Students could benefit from research agents, tutoring agents, archival agents, and communication agents. Research agents could constantly scour the infinite resources of the World Wide Web, reporting back to the student (or another agent) whenever any relative information is discovered. Tutoring agents would be highly customizable personal agents that grow with the student. Much like the aforementioned product brokering agents, these personal tutoring agents would remember the learning history of the student, keeping track of patterns of learning. This would enable the tutor to create a custom-built learning package that would cater directly to the child’s learning style.

Archival agents could track the student throughout the day, digesting every idea communicated (and piece of work created) by the student into a digital archive. In the CompILE environment, every interaction in the classroom would be recorded in digital format. These records would be readily available, produced quickly and quietly by the archival agents. If a student cannot remember exactly what a teacher said or wrote at the beginning of the day, he or she would no longer need to worry, with everything available on a permanent basis. Issues of storage would need to be considered, but a half-sufficient remedy could be the use of a single storage facility (accessible by all the students in the class) for all in-class communication. Communication agents representing each student could queue up for the teacher, and he or she could answer the questions in the order in which they were received. This would work much like the “spooling” process for a networked printer. A student could place the requests and continue working on some other task while his or her communication agent representative waited to be served.

The teacher could employ several intelligent teaching assistant agents to deal with the overload of student questions. Many students would likely repeat the same question, so one of these assistant agents could be created specifically to tackle this issue, essentially replicating the teacher’s original
answer for each of the students. Teachers could employ grading agents to assist them with grading exams, especially the objective evaluations. This portion of the grading process could be an automatic function of CompILE, with the teacher’s “grading” agents merely collecting the data from the system for the teacher’s records. Observational agents that track the student’s progress and individual learning styles could compare notes with the teacher’s lesson plan preparation agents (primarily used like research agents) to create customized lessons for the class that compensated for each student’s current position within the learning spectrum. It is even possible that these agents could continue working while the lesson is being presented, providing real-time diagnostics for the teacher, giving him or her the ability to change the speed or direction of the lesson on the fly!

As parents would have easy, secure access to CompILE, teachers could send conference requests via agents. As opposed to unanswered e-mails or phone calls that might have arrived at a time inconvenient to the parent, agents would be willing to wait indefinitely for a response, dutifully delivering it back to the teacher immediately upon receipt. Parents could also create agents to request progress reports on their children, or to create conferencing agents of their own, if they wished to speak with the teacher. Messenger agents could provide a way for parents to alert their children (and the teachers and administration) to any sudden change of plans that might arise on any particular day. This would be particularly helpful if the family has more than one child in the same school system, but different facilities (one in elementary school, one in middle school), or in entirely different systems. One agent could carry the message to all the different constituents involved.

In addition to progress report agents, parents could create agents that gathered their children’s homework assignments for the day (or week), creating a chart for easy time management, helping the children finish their homework efficiently and effectively. The homework agents could also serve as tutoring assistance agents, gathering information to assist the parent who might need a “refresher” on the fundamentals to be learned in the homework. These agents could report back to the teacher on the following day, giving a brief evaluation of the previous evening’s proceedings.

Administrators would have many uses for agents. In addition to agents already mentioned, the administration could employ agents to monitor teachers, creating periodical “evaluations” that indicate the progress of the students; one indication of the teacher’s ability to teach. Financial agents could help the administration cope with more political issues, such as which departments need more funding for equipment or textbooks, with research agents in turn scouring the resources of the Internet for the best options. Communication agents could serve the administration by contacting other school systems with similar demographic issues and comparing notes.

Researchers could employ the assistance of statistical agents that would perform data collection, sorting, and statistical analysis. Observation agents could continually observe the actions of different types of participants, searching for recognizable patterns of use of the CompILE system. As with administrators, communication agents could assist researchers by linking to the research taking place at other school systems with similar and different circumstances, performing trend comparisons and facilitating the collaboration of researchers on pertinent issues.

Design teams could rely heavily upon testing agents for the initial alpha testing of upgrades and new modules to be added to the CompILE application tool set, as the purpose of alpha testing is to find all possible ways to “break” software, and agents are the perfect digital drones for simulating endless combinations of user actions in an organized fashion. Role-playing agents could assist designers by demonstrating alternative
solutions in the scenario-based design approach. Documentation agents could be responsible for tracking and archiving each iteration within the design team’s process, allowing for appropriate reflection upon completion of a project.

Special care should be given when CompILE is implemented to prevent oversaturation of agents within the “cybersphere.” Most agents should be able to multitask. Or it is possible that the better solution would involve millions of minimalist agents designed to complete one simple task and dispose of themselves. In either case, the possibilities abound for agent use in education, and the main purpose of their existence should be to allow humans to perform their roles in a manner that embodies valid participation in the educational community. Additionally, the persistence of researchers and designers “on the scene” in schools should raise the comfort level of other participants (especially children) due to familiarity. Perhaps increased exposure to research and design processes for other participant roles would cause an increased general appreciation of these processes. More importantly, though, researchers would likely gain more honest insight into participant processes, especially on the qualitative level.

CompILE AND CULTURE

Ethnography is the process of identifying the customs, habits, and key differences between different races, cultures, and groups of people. Due to its small focus and personalized nature, ethnography is known for its rich descriptions providing historical depth and contextual perspective, tracing social processes within groups (Howard, 2002). The rich descriptions and contextual perspective can provide the foundation for the new form of education that accommodates comprehensive interactive learning environments. In terms of communication, ethnography is concerned with patterns and functions of communication, the nature and definition of speech community, the components of communicative competence, and linguistic and social universals and inequalities (Saville-Troike, 2003).

Patterns of communication exist on the societal, group, and individual level. The interconnectedness of these patterns largely affects the roles of individuals in society in general, as well as within an educational setting. If a person was unaware that the patterns differ from one group of people to the next, it would be easy to misunderstand the implications of a particular attempt at communication by someone from a different group or culture. The fact that our schools are becoming increasingly multicultural opens a floodgate of potential misconception.

The classroom can be viewed as a model for societal structure in the sense that it can serve as a microcosm of the communicative situations that exist in our society. Considering this, the need for more attentiveness to the context of the discourse makes sense, but is not always possible at the time the discourse takes place. Teachers have too many other things to think about, and the context of discourse cannot possibly always be their main concern. Teachers can have more time to explain the context of discourse to students, thanks to the streamlined efficiency of CompILE. Addressing this context “in the moment” can reinforce an open-mindedness in students that leads to a broader appreciation of all cultures.

In addition to a broader appreciation of the diversity of human cultures, CompILE can also foster an overall appreciation for “culture” from a more philosophical perspective, one that considers aesthetic beauty. Adler (1981) refers to the subjective side of beauty as “enjoyable beauty,” and the objectivity of beauty as “admirable beauty.” Gelernter (1998) takes the concept of admirable beauty and refines it into a more fitting term for the technological world: “machine beauty,” which can be considered a keen symmetry of simplicity and power. Conceptually, CompILE exemplifies this marriage of simplicity and power, providing access to so much information in such an organ-
nized, efficient way. However, its development falls under the shadow of the “beauty paradox”: (1) most computer technologists are oblivious to beauty; (2) the best are obsessed with it; (3) the public has a love-hate relationship with beauty in computing; (4) beauty is the most important quality that exists in the computer world, when all is said and done (Gelernter, 1998). The subjectivity of the love-hate relationship between users and computers is the biggest hindrance to the development of simple yet powerful applications such as CompILE. An application can be admirably beautiful to all its users, but some will not experience its enjoyable beauty, at first. A deeper appreciation for the admirable beauty can lead to a newfound appreciation and enjoyment.

FUTURE TRENDS: CompILE AND GLOBALIZATION

The globalization of our learning society will come about with the increased ubiquity of ICTs (including those yet to be realized), which is leading to the disappearance of the traditional “brick-and-mortar classroom space.” CompILE must be designed in a way to facilitate participants’ responses to this removal in a seamless, “beautiful” manner. A core of visual communication and literacy (Dondis, 1973), backed by a network of translation agents, will advance the intercultural community that would thrive in this global learning society, based entirely upon open access to information by all participants.

As mobile inhabitants of a global society, lifelong learners should not be held stationary by any sociotechnical system. A learner must be able to collaborate effectively, sharing the contents of his or her IKS with others using a sophisticated visualization interface. As we begin working across national and cultural boundaries with higher frequency, verbal communication in a “common language” might not resonate clearly, but recognition of patterns in other’s knowledge spaces, using basic elements of visual form could transcend this issue altogether. If we as a society were to become visually literate, would a common “visual language” lead to better understanding, communication, and cooperative learning? Perhaps this visual communication process could initially be augmented by translation agents, a specialized form of negotiation agents that augment the verbal communication between people who cannot find a common spoken language between them. These agents would be reliant upon the advent of highly accurate voice recognition technology, and instantaneous translation would be paramount to provide seamless assistance during conversation.

CONCLUSION

As a sociotechnical system, CompILE must manifest as a hybrid system in at least two ways. First, its users, in all forms of participation, would in fact be de facto designers. Second, CompILE would combine reality and virtual environments in a manner conducive to constructive learning practices, taking full advantages of the added dimensionality provided by this combination. Imagine a learner interested in Greek civilization, standing in modern-day Greece, able to augment his or her current visual field with images of classic Greek structures overlaid precisely upon those structures in their current form. Upon the request of the user, these augmenting images could animate, showing the degradation of the structures over time, highlighting access to information about specific historic events that had direct impact on the process. This is just one example of how CompILE could embody the careful integration of ICT as a tool fundamental to the learning process.

The agents of CompILE can help promote higher levels of information literacy amongst its users, which in time should lead to a more informed consumer society, which will provide
society with a more appropriate mindset to create community, embrace all cultures, and become a true learning society. The CompILE system, obviously, is a continuous work in progress, still in its infancy. Much more discussion must take place concerning theories and applications as a continuing predevelopment phase. It is precisely such a discussion, as well as the continued involvement in its design and development, that will empower users of CompILE to cross the digital divide.

REFERENCES


*This work was previously published in Social Information Technology: Connecting Society and Cultural Issues, edited by T. Kidd and I.L. Chen, pp. 348-362, copyright 2008 by Information Science Publishing (an imprint of IGI Global).*
Chapter 2.22
Problem Frames for Sociotechnical Systems

Jon G. Hall
The Open University, UK

Lucia Rapanotti
The Open University, UK

ABSTRACT

This chapter introduces Problem Frames as a framework for the analysis of sociotechnical problems. It summarizes the Problem Frames approach, its techniques and foundations, and demonstrates, through theory and examples, how it can be applied to simple sociotechnical systems. The chapter continues with the description of an extended Problem Frame framework that allows the treatment of more general sociotechnical problems. This extension covers social components of a system — individuals, groups or organisations — bringing them within the remit of the design activity. The aim of the chapter is to make the Problem Frames framework more accessible to the software practitioner, especially those involved in the analysis of sociotechnical problems, as these problems have so far received only scant coverage in the Problem Frames literature.

INTRODUCTION

By sociotechnical system we mean a collection of interacting components in which some of the components are people and some are technological. In this chapter we focus on the requirements analysis of sociotechnical systems in which some of the technological subsystems are computer-based, these systems forming the largest part of modern software design problems.

More precisely, there are two (not necessarily disjoint) sub-classes of sociotechnical systems that we will treat in this chapter. The first sub-class contains those systems in which existing components or sub-systems (that is, domains) are to be allowed, through software, to interact. An example from this first class might be the problem of designing software for the operator of heavy machinery. The larger second class contains those systems for which software, a user interface, and user instruction are to be designed to enable a new process or service. An example
of this second class might be the development of a new customer call centre.

The use of Problem Frames (PFs) underpins our requirements analysis process. As described in Jackson (1998), PFs are a concretization of the ideas of Michael Jackson and others in the separation of machine and its environment’s descriptions. This separation is generally accepted as being a useful principle for requirements analysis. We will have cause, later in the chapter, in dealing with a more general class of sociotechnical problems, to further detail this separation, but nothing we do compromises its fundamental status.

The usual representation of the separation of machine and environment descriptions is as the “two ellipse” model, illustrated in Figure 1. In that figure world knowledge $W$ is a description of the relevant environment; $R$ is the statement of requirements; $S$ is the specification that mediates between environment and machine; $M$ is the description of the machine; and $P$ is the program that, on machine $M$, implements the specification $S$. The role of $W$ is to bridge the gap between specification $S$ and requirements $R$. More formally (Gunter, Gunter, Jackson, & Zave, 2000; Hall & Rapanotti, 2003; Zave & Jackson, 1997), $W,S \models R$.

One of the aims of the PF framework is to identify basic classes of problems that recur throughout software development. Each such class should be captured by a problem frame that provides a characterization for the problem class. Sociotechnical systems are an important class of problems and so should be representable within the PF framework, possibly with their own (collection of) problem frame(s).

In a fundamental sense, of course, the PF framework already deals with sociotechnical systems: problem frames are an attempt to allow customer and developer to come together to match real-world problems and technological solutions. There are many examples in Jackson (2001) as to how this relationship can be facilitated using problem frames.

We observe, however, that the application of problem frames to particular sociotechnical problems remains under-explored. Currently some discussion of HCI appears in Jackson (2001), and some analysis appears in Jackson (1998), but otherwise there is little in-depth coverage of how to apply problem frames in this context. In this chapter we show, in some detail, how problem frames can be applied to sociotechnical systems.

Our development is threefold. We first show how the problem of representing interaction with (and not just control of) technology can be represented within the PF framework. To do this we introduce two new basic problem frames, the User Interaction Frame and the User Commanded Behaviour Frame, each dealing with the class of user-interaction problems.

Secondly we show how architectural artifacts can be used to guide the analysis of sociotechnical problems. To do this we discuss the notion of an Architectural Frame (AFrame for short), a new PF artifact that can be used to guide problem decomposition in the light of particular solution expertise as might, for instance, exist in a software development company. As an exemplar of AFrames and their use, we define and apply an AFrame corresponding to the Model View Controller (MVC) architectural style (Bass, Clements, & Kazman, 1998).

Lastly we adapt the PF framework to meet the needs of representing the problems of more

---

**Figure 1. The requirements analysis model**

![Figure 1. The requirements analysis model](image)
Problem Frames for Sociotechnical Systems

complex sociotechnical systems, including those in which, as well as user-machine interaction, user training needs to be addressed. This causes us to consider a reification of the two-ellipse model into three ellipses to represent machine, environment, and user descriptions. Consequently, by interpreting this third ellipse in the PF framework, we discover a new type of PF domain—the knowledge domain—to represent a user’s knowledge and their “instruction” needs, and argue that this leads to a more general PF framework for sociotechnical systems.

Chapter Overview

In the major part of this chapter we will use a well-known chemical reactor problem (Dieste & Silva, 2000; Leveson, 1986) as an example for illustration of techniques. Later we briefly describe the design of a “cold-calling” system. The chemical reactor is a sociotechnical system and is representative of the class of operator-controlled safety-(and mission-)critical systems. A schematic for the chemical reactor hardware appears in Figure 2.

A statement of the problem is as follows:

A computer system is required to control the safe and efficient operation of the catalyst unit and cooling system of a chemical reactor. The system should allow an operator to issue commands for activating or deactivating the catalyst unit, and to monitor outputs. Based on the operator’s commands, the system should instruct the unit accordingly and regulate the flow of cooling water. Attached to the system is a gearbox: whenever the oil level in the gearbox is low, the system should alert the operator and halt execution.

The chapter is organized as follows. The next section develops the problem frame representation of the chemical reactor problem and uses this to recall the basis of problem representation in the PF framework. The section on problem classification provides a small taxonomy of problem classes, including those of relevance to sociotechnical systems. The next section addresses problem decomposition, both in the classical PF framework and through AFrames. The section on requirements analysis models for sociotechnical systems details our separation into three of the various domain descriptions and uses this to motivate a new type of PF domain, the knowledge domain. The final section concludes the chapter.

PROBLEM REPRESENTATION

We first consider the PF representation of the chemical reactor problem. Within the PF framework problems are represented through problem diagrams. A problem diagram defines the shape

Figure 2. The chemical reactor schematic (adapted from Dieste & Silva, 2000)
of a problem: it records the characteristic descriptions and interconnections of the parts (or domains) of the world the problem affects; it places the requirements in proper relationship to the problem components; it allows a record of concerns and difficulties that may arise in finding its solution.

For the chemical reactor, there are a number of domains, including those that appear in the schematic of Figure 2. Also the operator will play an important role in issuing commands to control the catalyst and cooling systems. Placing all of these domains in their correct relationship to each other leads to the problem diagram shown in Figure 3.

The components are:

- **Operation machine**: The machine domain, that is, the software system and its underlying hardware. The focus of the problem is to build the Operation machine.

- **Other boxes** (Cooling System, Catalyst, and so forth): Given domains representing parts of the world that are relevant to the problem.

- **Shared phenomena**: The ways that domains communicate. These can include events, entities, operations, and state information. In Figure 3, for example, the connection between the Operation machine and the Cooling System is annotated by a set $e$, containing the events $\text{increase}_{\text{water}}$ and $\text{decrease}_{\text{water}}$, and a set $f$, containing the phenomenon $\text{water}_{\text{level}}$. Phenomena in $e$ are controlled by the Operation machine; this is indicated by an abbreviation of the domain name followed by $!$, that is, $\text{OM}!$. Similarly the phenomenon in $f$ is controlled by the Cooling System, indicated by the $\text{CS}!$.

- The dotted oval. **Safe and efficient operation**: The requirement, that is, what has to be true

---

**Figure 3. The chemical reactor problem diagram**

```
- $a: \{\text{open}_{\text{catalyst}}, \text{close}_{\text{catalyst}}\}$
- $b: \{\text{catalyst}_{\text{status}}, \text{water}_{\text{level}}\}$
- $c: \{\text{open}_{\text{catalyst}}, \text{close}_{\text{catalyst}}\}$
- $d: \{\text{is}_{\text{open}}, \text{is}_{\text{closed}}\}$
- $e: \{\text{increase}_{\text{water}}, \text{decrease}_{\text{water}}\}$
- $f: \{\text{water}_{\text{level}}\}$
- $g: \{\text{oil}_{\text{level}}\}$
```
of the world for the (operation) machine to be a solution to the problem.

- In the connections between the requirement and the domains, a dotted line indicates that the phenomena are referenced (that is, an object of) the requirement, while a dotted arrow indicates that the phenomena are constrained (that is, a subject for the requirement). In Figure 3, for instance, the oil level in the gear box is referenced while the cooling system’s water level is constrained.

- Phenomena at the requirement interface (for example, those of sets \( f \) or \( d \)) can be, and usually are, distinct from those at the machine domain interface (for example, those of sets \( c \) and \( e \)). The former are called requirement phenomena; the latter, specification phenomena. The intuition behind this distinction is that the requirement is expressed in terms of elements of the problem, while the specification (that is, what describes a machine domain) is expressed in terms of elements of the solution.

Other artifacts that are not represented on the problem diagram but are related to it are domain and requirement descriptions. Such descriptions are essential to the analysis of a problem and address relevant characteristics and behaviours of all given domains, the machine domain, and the requirement.

An important distinction in the PF framework is that of separating two types of descriptions: indicative and optative. Indicative descriptions are those that describe how things are; optative descriptions describe how things should be. In this sense, in a problem diagram, given domain descriptions are indicative, while requirement and machine descriptions are optative. In other words, things that have to do with the problem domain are given, while things that have to do with the solution domain can be chosen. For instance, in the chemical reaction problem indicative descriptions of the Catalyst, Cooling System, and Gear Box should include characteristics of the domains that are of interest in the specification of the machine, say, the mechanics of opening the catalyst, or of changing the water level in the cooling system, or the oil level in the gear box. On the other hand the requirement should express some constraints on the status and operations of those domains that, when satisfied, result in their safe and efficient operation. Finally the machine specification should describe how we would like the control system to behave and interact with the given domains so that the requirements are met.

Indeed not all domains share the same characteristics. There is a clear difference between a cooling system and an operator. In a cooling system there exist some predictable causal relationships among its phenomena. For instance, it could be described as a state machine, with a set of clearly identified states and predictable transitions between them. A domain with such characteristics is known as a causal domain. On the other hand, an operator’s phenomena lack such predictable causal relationships. We can describe the actions an operator should be allowed to do but cannot guarantee that they will be executed in any particular order, or not at all, or that some other (unpredicted) actions will not be executed instead. A domain with such characteristics is known as biddable.

The distinction between causal and biddable domains is an important one, as it has ramifications for the type of descriptions we can provide and the assumptions we can make in discharging proof obligations during the analysis of a problem, as we will see in the following sections.

Of course there exist other types of domains, each with its own characteristics. An exhaustive domain classification is beyond the scope of this chapter and can be found in, for example, Jackson (1995), Jackson (1998) and Jackson (2001).
PROBLEM CLASSIFICATION

One of the intentions of the PF framework is to classify problems. An initial problem classification is given in Jackson (2001). Therein are identified five basic problem frames. They are basic because they represent relatively simple, recurrent problems in software development. In describing the basis of problem frames, the intention is not to be exhaustive in its classification of problems. Indeed there are other frames by Jackson (see, for example, Jackson, 1998) that do not make it into that classification.

In this section we present a simple taxonomic development, beginning with the simplest of all problem frames and adding domain types (modulo topology). As each type has different characteristics, the resulting problem frames represent different classes of problems. In doing so we introduce a new basic problem frame, the User Interaction Frame, which is novel in the problem class it represents within the PF framework (but not, of course, within software engineering). The taxonomy is summarized in Figure 4.

Programs

The simplest form of problem representable in problem frames is that of producing a program from a given specification. This is illustrated in Figure 5. Although this is a sub-problem of all software engineering problems, it is not a very interesting problem class to be analyzed using problem frames: nothing exists outside the machine. Other techniques, such as JSD (Jackson, 1997) or design patterns (Gamma, Helm, Johnson, & Vlissides, 1995), are probably more appropriate to tackle this problem class. Indeed the PF framework does not include a problem frame for this class of problems.

Embedded Controllers

For a problem class to be purposefully analyzed in PFs, some given domain(s) of interest must exist outside the machine. An interesting problem class is identified in Jackson (2001) by introducing a single causal domain. The problem frame is known as the Required Behaviour Frame, and its characterizing problem is that of building a machine that controls the behaviour of some part of the physical world so that it satisfies certain conditions. Some software engineers may find it easy to identify this problem with that of building an embedded controller (although it does apply also to more general software problems).

The Required Behaviour Frame is illustrated in Figure 6. The frame has a topology that is captured by a frame diagram (that of the illustration).

The frame diagram resembles a problem diagram, but it also includes some further annotation.
This provides an indication of the characteristics of the domains and phenomena involved in problems of the class. For the Required Behaviour Frame:

- The **Controlled domain** is causal (the C annotation in the figure). Its phenomena are also causal (indicated by C on the arcs) — they are directly caused or controlled by a domain and may cause other phenomena in turn.
- The **Control machine** has access to causal phenomena of the **Controlled domain** (in C2) and controls another set of phenomena, which are also shared with the **Controlled domain** (in C1). Intuitively phenomena in C1 are used by the machine to control the domain, while phenomena in C2 are used to obtain information and feedback on the functioning and state of the domain.
- The requirement, **Required behaviour**, is expressed in terms of a set (C3) of causal phenomena of the **Controlled domain**.

When a problem of a particular class is identified, it can be analyzed through the instantiation of the corresponding frame diagram. The instantiation is a process of matching problem and frame’s domains and their types, as well as problem and frame’s phenomena types. The result of the instantiation is a problem diagram, which has the same topology of the frame diagram but with domains and phenomena grounded in the particular problem.

Let us return to the chemical reactor problem and consider the problem of regulating the water level in isolation. This can be regarded as a required behaviour problem (Figure 7) with some safety requirement on the water level.

For a problem to be fully analyzed, the instantiation of a problem frame is only the first step of the process. Suitable domain and requirement descriptions (see the section on problem representation) need to be provided and the frame concern needs to be addressed. The frame concern is an overall correctness argument, common to all the problems of the class. It is the argument that must convince you, and your customer, that the specified machine will produce the required behaviour once combined with the properties of the given domains. Each problem frame comes with a particular concern whose structure depends on the nature of the class problem. For

---

**Figure 6. Required Behaviour Frame**

![Required Behaviour Frame](image)

**Figure 7. Regulating the water level in the cooling system as a required behaviour problem**

![Water Level Regulation Diagram](image)

```
e:{increase_water,decrease_water}
f:{water_level}
```
the Required Behaviour Frame, the argument is outlined in Figure 8.

**User Interaction**

Another interesting class of problems can be obtained by including a single biddable domain outside the machine, which represents the user of the system. We call the resulting problem frame the *User Interaction Frame*, and its characterizing problem is that of building a machine that enforces some rule-based interaction with the user. The frame diagram for the User Interaction Frame is given in Figure 9.

The *Interaction machine* is the machine to be built. The *User* is a biddable domain representing the user who wants to interact with the machine. The requirement gives the rules that establish legal user/machine interactions.

The manifestation of the user/machine interaction is through exchanges of causal phenomena (in $C_1$) controlled by the user and symbolic phenomena (in $Y_1$) controlled by the machine. Intuitively the user issues commands in $C_1$ and the machine provides feedback through $Y_1$. The interaction rules specify the legal correspondence of user and machine phenomena.

In the chemical reactor problem we can isolate the operator/machine interaction as a user interaction problem as illustrated in Figure 10, where the requirement establishes some rules on the relationship between operator’s commands and system feedback, say, that an *open_catalyst* command cannot be issued when the *water_level* value is below a set threshold. Note that this would be the perspective of a User Interface (UI) designer, whose main concern is the user interacting with a black box system. Indeed, in the wider problem analysis of the chemical reactor problem, machine responses to user commands depend on a faithful representation of the internal state of, say, the catalyst or the cooling system.
Figure 11 illustrates the frame concern for the User Interaction Frame.

**User Commanded Behaviour**

The Required Behaviour and the User Interaction Frames are representative of relatively simple problems, albeit often recurring in software development. It is possible, and indeed likely, that other interesting problem classes could be identified by considering single given domains of some other type (see discussion at the end of the section on problem representation). However we do not pursue this any further here. We look, instead, at what happens when there are two given domains outside the machine.

As for the single domain case, other interesting classes of problems emerge. In fact the remaining four basic problem frames introduced in Jackson (2001) are all of this form.

If we add a biddable domain to the Requirement Behaviour Frame, we obtain a User Commanded Behaviour Frame, which is illustrated in Figure 12. Its characterizing problem is that of building a machine that will accept the user’s commands, impose control on some part of the physical world accordingly, and provide suitable feedback to the user. Jackson (2001) introduces a subclass of this frame, the Commanded Behaviour Frame, which does not require the user to receive any feedback.

---

**Figure 10. Operator/system interaction as an instance of the User Interaction Frame**

\[
\text{Interaction} \quad \begin{array}{c}
\text{machine} \\
\text{IM/b} \\
\text{OP/a}
\end{array} \quad \begin{array}{c}
\text{Operator} \\
\text{a,b}
\end{array} \quad \text{Rules of interaction}
\]

\(a: \{\text{open\_catalyst, close\_catalyst}\} \quad b: \{\text{catalyst\_status, water\_level}\}\)

**Figure 11. Frame concern for the User Interaction Frame**

1. Given this set of machine phenomena, when the user causes this phenomena (it may or may not be sensible or viable)...  
2. if sensible or viable the machine will accept it...  
3. resulting in this set of machine phenomena...  
4. thus achieving the required interaction in every case.
In the chemical reactor problem, we can apply the User Commanded Behaviour Frame to analyze how the catalyst is controlled by the operator. The corresponding problem diagram is given in Figure 13.

A possible description of the interaction rules could be as follows. The machine shall allow the user to control the catalyst under the following constraints:

1. *catalyst_status* is a faithful representation of the state of the catalyst
2. the initial state of the catalyst is *catalyst_closed*
3. possible user commands are *open_catalyst* or *close_catalyst*
4. state transitions are represented in Figure 14.

The frame concern for the User Commanded Behaviour Frame is given in Figure 15. From the figure you will notice that the argument has two parts: satisfying the required behaviour of the domain (from 1 to 4) and providing suitable feedback to the user (5 and 6).

**PROBLEM DECOMPOSITION**

Most real problems are too complex to fit basic problem frames. They require, rather, the structur-
Problem Frames for Sociotechnical Systems

In this section we discuss two ways of decomposing problems within the PF framework. The first, classical decomposition, proceeds through sub-problem identification and problem frame instantiation. The second, our novel approach, combines sub-problem identification with guided architectural decomposition using AFrames.

Classical Decomposition

In classical PF decomposition a problem is decomposed into simpler constituent sub-problems that
Problem Frames for Sociotechnical Systems

can then be analysed separately. If necessary each sub-problem can be decomposed further, and so forth, until only very simple sub-problems remain. Decomposition proceeds through the identification of sub-problems that fit a recognised problem class and the instantiation of the corresponding problem frame. We illustrate the process on the chemical reactor problem.

There are three sub-problems:

1. a user-commanded behaviour problem, for the operator to control the catalyst;
2. a required behaviour problem, for regulating the water flow in the cooling system; and
3. a sub-problem to issue a warning (and halt the system) when there is an oil leak in the gearbox.

Addressing sub-problems 1 and 2 means instantiating the corresponding problem frames to derive problem diagrams for each sub-problem. These were depicted in Figures 13 and 7, respectively.

The third sub-problem has no standard problem frame to represent it. The closest fit would be the Information Display Frame (Jackson, 2001), but this requires a decision on how the alarm will be raised. Here we have made an arbitrary choice of introducing a Bell domain, and assume that it will

Figure 16. Raising the alarm as an instance of the information display problem

![Diagram of raising the alarm as an instance of the information display problem]

h: \{ring_bell\}  i: \{bell_ringing\}  g: \{oil_level\}

Figure 17. Further decomposition of the user Commanded Behaviour sub-problem

![Diagram of further decomposition of the user Commanded Behaviour sub-problem]

a: \{open_catalyst, close_catalyst\}  b: \{catalyst_status\}  c: \{open_catalyst, close_catalyst\}  d: \{is_open, is_closed\}
ring when the oil level in the gearbox is below a certain threshold. The resulting sub-problem diagram is shown in Figure 16.

Finally we already know from the simple taxonomy in the third section that sub-problem one can be decomposed further, resulting in a required behaviour and a user interaction sub-problem. These are shown in Figure 17.

**AFrames**

AFrame decomposition complements classical decomposition in providing guidance and decomposition and recomposition rules. The rational behind AFrames is the recognition that solution structures can be usefully employed to inform problem analysis.

AFrames characterise the combination of a problem class and an architectural class. An AFrame should be regarded as a problem frame for which a “standard” sub-problem decomposition (that implied by an architecture or architectural style) exists. AFrames are a practical tool for sub-problem decomposition that allow the PF practitioner to separate and address, in a systematic fashion, the concerns arising from the intertwining of problems and solutions, as has been observed to take place in industrial software development (Nuseibeh, 2001). Further motivation for, and other examples of, AFrames can be found in Rapanotti, Hall, Jackson, & Nuseibeh (2004).

The MVC (short for Model-View-Controller) (see, for example, Bass et al., 1998) is a way of structuring a software solution into three parts — a model, a view, and a controller — to separate and handle concerns related, respectively, to the modeling of a domain of interest, the visual feedback to the user, and the user input. The controller interprets user inputs and maps them into commands to the model to effect the appropriate change. The model manages one or more data elements, responds to queries about its state, and responds to instructions to change state. The view is responsible for feedback on the model’s state to the user. Standard communication patterns (for example, the Observer pattern (Gamma et al., 1995)) apply between the MVC parts.

Here we introduce the MVC AFrame as applied to the User Commanded Behaviour Frame. This represents the class of user commanded behaviour problems for which an MVC solution is to be provided.

The intention of using the MVC in the solution space is recorded through an annotation of the machine as illustrated in Figure 18. Guidance on decomposition is in the form of decomposition templates, which are applied to obtain sub-problem diagrams. The decomposition templates for the MVC AFrame are given in Figure 19. It can be seen from the figure that the original problem is decomposable into two sub-problems, whose

---

*Figure 18. MVC annotation of the User Commanded Behaviour Frame*
machine domains are the View and Controller machines (in the MVC sense). Also a Model domain is introduced that represents an abstraction of the real-world domain to be controlled. This is a designed domain (Jackson, 2001), that is, one that we have the freedom to design, as it will reside inside the solution machine. The resulting sub-problems are then: that of building a View machine to display the Model’s representation of the state of the controlled domain and that of building a Controller machine that acts on the Model, which will pass on the commands to the controlled domain. In PF terms the Model acts as a connection domain between the real-world domain and presentation and control subsystems.

The application of the MVC AFrame to the sub-problem of controlling the catalyst (see Figure 13) results in the decomposition of Figure 20.

We see at least two strengths of AFrames. The first is that they suggest how a problem would need to be restructured for a particular solution form. For instance, in the MVC case, that an abstract model of the catalyst needs to be produced (or, for that matter, a connection domain between Operator and Gearbox — a Bell — would be needed).

The second is that they help the recomposition of sub-problem solutions into the original problem. Recomposition is facilitated by the fact that AFrame decomposition is regularized through the application of the AFrame templates. For the MVC this is through the identification of the links among its architectural elements. The recomposition diagram for the MVC AFrame is illustrated in Figure 21 and its frame concern in Figure 22. In Figure 21 the recomposition of the model, view, and controller domains follows the MVC architectural rules.
Problem Frames for Sociotechnical Systems

Figure 20. MVC decomposition of the user commanded behaviour sub-problem

![MVC decomposition diagram]

Figure 21. MVC recomposition

![MVC recomposition diagram]

A REQUIREMENTS ANALYSIS MODEL FOR SOCIOTECHNICAL SYSTEMS

The consideration of more sophisticated human-machine relationships is our next concern. To be specific we now wish to look at users’ behaviour as being the subject of requirements statements, admitting that users are the source of much of the flexibility in sociotechnical systems. In short we wish to allow the design of human instruction to be the subject of the requirements engineering process addressed through problem frames alongside that of the program.

Paraphrasing this we might say that the human, as well as the machine, is to be the subject of
Problem Frames for Sociotechnical Systems

Figure 22. Discharging the correctness argument in MVC recomposition

1. Given a choice of commands in the current state, when the user issues this command (it may or may not be sensible)...
2. If sensible or viable, the machine will cause these events...
3. Resulting in this state or behaviour...
4. Which satisfies the requirement...
5. And which the machine will relate to the user...
6. Thus satisfying the requirement in every case.

optative descriptions. Foundationally this means the separation of the description of the world from that of the human who is the subject of the design. This leads to the reification of the original ellipse model shown in Figure 23. In it we have three ellipses – that for the Human $H$ with knowledge $K$, that for the Machine $M$ with program $P$, and that for the remaining Environment $W$ with requirements $R$. Of course, just as machines outside the design process have indicative descriptions in $W$, so do humans.

With the introduction of the human $H$, we identify and separate two new areas of interest, which now form explicit foci for design:

- the specification $UI$, anonymous in the $S$ region in the original model, which determines the Human-Machine interface; and
- the specification $I$, missing from the original model, which determines the knowledge and behaviour that is expected of the human as a component of the sociotechnical system.

As in the original model the description $W$ has the role of bridging the gaps between the requirement $R$ and the specification $S$, in our extension $W$ has the role of bridging the gaps between the requirement $R$ and the instruction $I$, human-machine interface $UI$ and specification $S$ together. More precisely we assert that $S$, $I$, $UI$, and $W$ must be sufficient to guarantee that the requirements of the sociotechnical system are satisfied. More formally:
Problem Frames for Sociotechnical Systems

A Problem Frame Interpretation

In the PFs framework the machine domain represents the machine for which the specification $S$ must be designed. By analogy a new domain type will be required to represent the human for which the instruction $I$ has to be designed. To this end we introduce into the PF framework the notion of a knowledge domain to represent that domain. In a problem diagram a knowledge domain should be represented as a domain box with a double bar on the right-hand side (to distinguish it from the machine domain).

The most general form of a sociotechnical problem, as a problem diagram, is shown in Figure 24. In the figure both Knowledge and Machine domains are subjects of design, as are their shared user interface phenomena.

An example of how a real-world sociotechnical problem could be treated in the new model is that of designing a “cold-calling” system to allow an interviewee to be telephoned and for their responses to certain questions, asked by an interviewer, to be recorded in a database for future analysis. The problem is to design both the technical subsystem (the machine domain) and the instructions that guide the interaction of the interviewer (the knowledge domain) with the interviewee. The interviewee sits firmly in the environment (as a biddable, indicative domain). The interaction with the database is a machine task.

The outcome of the design process, in addition to the specification for the technical subsystem, might be a script for the interviewer and the human-machine interface as used by the interviewer.
The problem diagram for this example is outlined in Figure 25.

CONCLUSION

In their classical form problem frames happily represent interactions between a user and a machine, as might be characteristic of simple sociotechnical systems. In this chapter we have presented an enrichment of the PF framework to allow the representation and analysis of more complex sociotechnical systems. To do this we have introduced two new problem frames, the User Interaction and User Commanded Behaviour Frames. Although not exhaustive in their treatment of socio-technological interaction problems, they hopefully will provide a sound basis for a richer taxonomy of user interaction within the PF framework.

One of the as-yet under-developed areas within the PF framework is the treatment of problem decomposition, in particular from the perspective of how to do it in practice. We are currently exploring the development and use of A Frames. An A Frame offers guidance in problem decomposition on the basis of solution space structures. In this chapter we have shown how basic sociotechnical interaction problems can be decomposed when the target architecture is to be the MVC style.

Although both these enrichments are new in the PF framework, they do not move outside of its original conceptual basis in the two-ellipse model of requirements analysis. In contrast we have seen in this chapter that the development of general sociotechnical systems raises challenges for the PF framework. We have suggested solutions to these challenges in the reification of the two-ellipse model to a three-ellipse version, in which social sub-systems – individuals, groups, and organisations – can also be considered as the focus of the design process. With the introduction of the knowledge domain, the manifestation of this extension in problem frames, we aim to bring the sociotechnical system problems under the design remit of the PF framework.

ACKNOWLEDGMENT

We acknowledge the kind support of our colleagues, especially Michael Jackson and Bashar Nuseibeh in the Department of Computing at the Open University.

REFERENCES


This work was previously published in *Requirements Engineering for Sociotechnical Systems*, edited by J. L. Mate & A. Silva, pp. 318-339, copyright 2005 by Information Science Publishing (an imprint of IGI Global).
Chapter 2.23

Integrating Semantic Knowledge with Web Usage Mining for Personalization

Honghua Dai
DePaul University, USA

Bamshad Mobasher
DePaul University, USA

ABSTRACT

Web usage mining has been used effectively as an approach to automatic personalization and as a way to overcome deficiencies of traditional approaches such as collaborative filtering. Despite their success, such systems, as in more traditional ones, do not take into account the semantic knowledge about the underlying domain. Without such semantic knowledge, personalization systems cannot recommend different types of complex objects based on their underlying properties and attributes. Nor can these systems possess the ability to automatically explain or reason about the user models or user recommendations. The integration of semantic knowledge is, in fact, the primary challenge for the next generation of personalization systems. In this chapter we provide an overview of approaches for incorporating semantic knowledge into Web usage mining and personalization processes. In particular, we discuss the issues and requirements for successful integration of semantic knowledge from different sources, such as the content and the structure of Web sites for personalization. Finally, we present a general framework for fully integrating domain ontologies with Web usage mining and personalization processes at different stages, including the preprocessing and pattern discovery phases, as well as in the final stage where the discovered patterns are used for personalization.

INTRODUCTION

With the continued growth and proliferation of e-commerce, Web services, and Web-based information systems, personalization has emerged as a critical application that is essential to the success of a Website. It is now common for Web users to encounter sites that provide dynamic recommendations for products and services, tar-
geted banner advertising, and individualized link selections. Indeed, nowhere is this phenomenon more apparent as in the business-to-consumer e-commerce arena. The reason is that, in today’s highly competitive e-commerce environment, the success of a site often depends on the site’s ability to retain visitors and turn casual browsers into potential customers. Automatic personalization and recommender system technologies have become critical tools, precisely because they help engage visitors at a deeper and more intimate level by tailoring the site’s interaction with a visitor to her needs and interests.

Web personalization can be defined as any action that tailors the Web experience to a particular user, or a set of users (Mobasher, Cooley & Srivastava, 2000a). The experience can be something as casual as browsing a Website or as (economically) significant as trading stocks or purchasing a car. Principal elements of Web personalization include modeling of Web objects (pages, etc.) and subjects (users), categorization of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization. The actions can range from simply making the presentation more pleasing to anticipating the needs of a user and providing customized information.

Traditional approaches to personalization have included both content-based and user-based techniques. Content-based techniques use personal profiles of users and recommend other items or pages based on their content similarity to the items or pages that are in the user’s profile. The underlying mechanism in these systems is usually the comparison of sets of keywords representing pages or item descriptions. Examples of such systems include Letizia (Lieberman, 1995) and WebWatcher (Joachims, Freitag & Mitchell, 1997). While these systems perform well from the perspective of the end user who is searching the Web for information, they are less useful in e-commerce applications, partly due to the lack of server-side control by site owners, and partly because techniques based on content similarity alone may miss other types of semantic relationships among objects (for example, the associations among products or services that are semantically different, but are often used together).

User-based techniques for personalization, on the other hand, primarily focus on the similarities among users rather than item-based similarities. The most widely used technology user-based personalization is collaborative filtering (CF) (Herlocker, Konstan, Borchers & Riedl, 1999). Given a target user’s record of activity or preferences, CF-based techniques compare that record with the historical records of other users in order to find the users with similar interests. This is the so-called neighborhood of the current user. The mapping of a visitor record to its neighborhood could be based on similarity in ratings of items, access to similar content or pages, or purchase of similar items. The identified neighborhood is then used to recommend items not already accessed or purchased by the active user. The advantage of this approach over purely content-based approaches that rely on content similarity in item-to-item comparisons is that it can capture “pragmatic” relationships among items based on their intended use or based on similar tastes of the users.

The CF-based techniques, however, suffer from some well-known limitations (Sarwar, Karypis, Konstan & Riedl, 2000). For the most part these limitations are related to the scalability and efficiency of the underlying algorithms, which requires real-time computation in both the neighborhood formation and the recommendation phases. The effectiveness and scalability of collaborative filtering can be dramatically enhanced by the application of Web usage mining techniques.

In general, Web mining can be characterized as the application of data mining to the content, structure, and usage of Web resources (Cooley, Mobasher & Srivastava, 1997; Srivastava, Cooley, Deshpande & Tan, 2000). The goal of Web min-
Integrating Semantic Knowledge with Web Usage Mining for Personalization

The goal of Web usage mining, in particular, is to capture and model Web user behavioral patterns. The discovery of such patterns from the enormous amount of data generated by Web and application servers has found a number of important applications. Among these applications are systems to evaluate the effectiveness of a site in meeting user expectations (Spiliopoulou, 2000), techniques for dynamic load balancing and optimization of Web servers for better and more efficient user access (Pitkow & Pirolli, 1999), and applications for dynamically restructuring or customizing a site based on users’ predicted needs and interests (Perkowitz & Etzioni, 1998).

More recently, Web usage mining techniques have been proposed as another user-based approach to personalization that alleviates some of the problems associated with collaborative filtering (Mobasher et al., 2000a). In particular, Web usage mining has been used to improve the scalability of personalization systems based on traditional CF-based techniques (Mobasher, Dai, Luo & Nakagawa, 2001, 2002).

However, the pure usage-based approach to personalization has an important drawback: the recommendation process relies on the existing user transaction data, and thus items or pages added to a site recently cannot be recommended. This is generally referred to as the “new item problem”. A common approach to resolving this problem in collaborative filtering has been to integrate content characteristics of pages with the user ratings or judgments (Claypool et al., 1999; Pazzani, 1999). Generally, in these approaches, keywords are extracted from the content on the Website and are used to either index pages by content or classify pages into various content categories. In the context of personalization, this approach would allow the system to recommend pages to a user, not only based on similar users, but also (or alternatively) based on the content similarity of these pages to the pages the user has already visited.

Keyword-based approaches, however, are incapable of capturing more complex relationships among objects at a deeper semantic level based on the inherent properties associated with these objects. For example, potentially valuable relational structures among objects such as relationships between movies, directors, and actors, or between students, courses, and instructors, may be missed if one can only rely on the description of these entities using sets of keywords. To be able to recommend different types of complex objects using their underlying properties and attributes, the system must be able to rely on the characterization of user segments and objects, not just based on keywords, but at a deeper semantic level using the domain ontologies for the objects. For instance, a traditional personalization system on a university Website might recommend courses in Java to a student, simply because that student has previously taken or shown interest in Java courses. On the other hand, a system that has knowledge of the underlying domain ontology might recognize that the student should first satisfy the prerequisite requirements for a recommended course, or be able to recommend the best instructor for a Java course, and so on.

An ontology provides a set of well-founded constructs that define significant concepts and their semantic relationships. An example of an ontology is a relational schema for a database involving multiple tables and foreign keys semantically connecting these relations. Such constructs can be leveraged to build meaningful higher-level knowledge in a particular domain. Domain ontologies for a Website usually include concepts, subsumption relations between concepts (concept hierarchies), and other relations among concepts that exist in the domain that the Web site represents. For example, the domain ontologies of a movie Website usually include concepts such as “movie,” “actor,” “director,”
“theater,” and so forth. The genre hierarchy can be used to represent different categories of movie concepts. Typical relations in this domain may include “Starring” (between actors and movies), “Directing,” “Playing” (between theaters and movies), and so forth.

The ontology of a Website can be constructed by extracting relevant concepts and relations from the content and structure of the site, through machine learning and Web mining techniques. But, in addition to concepts and relations that can be acquired from Web content and structure information, we are also interested in usage-related concepts and relations in a Website. For instance, in an e-commerce Website, we may be interested in the relations between users and objects that define different types of online activity, such as browsing, searching, registering, buying, and bidding. The integration of such usage-based relations with ontological information representing the underlying concepts and attributes embedded in a site allows for more effective knowledge discovery, as well as better characterization and interpretation of the discovered patterns.

In the context of Web personalization and recommender systems, the use of semantic knowledge can lead to deeper interaction of the visitors or customers with the site. Integration of domain knowledge allows such systems to infer additional useful recommendations for users based on more fine-grained characteristics of the objects being recommended, and provides the capability to explain and reason about user actions.

In this chapter we present an overview of the issues related to and requirements for successfully integrating semantic knowledge in the Web usage mining and personalization processes. We begin by providing some general background on the use of semantic knowledge and ontologies in Web mining, as well as an overview of personalization based on Web usage mining. We then discuss how the content and the structure of the site can be leveraged to transform raw usage data into semantically-enhanced transactions that can be used for semantic Web usage mining and personalization. Finally, we present a framework for more systematically integrating full-fledged domain ontologies in the personalization process.

**BACKGROUND**

**Semantic Web Mining**

Web mining is the process of discovering and extracting useful knowledge from the content, usage, and structure of one or more Web sites. Semantic Web mining (Berendt, Hotho & Stumme, 2002) involves the integration of domain knowledge into the Web mining process.

For the most part the research in semantic Web mining has been focused in application areas such as Web content and structure mining. In this section, we provide a brief overview and some examples of related work in this area. Few studies have focused on the use of domain knowledge in Web usage mining. Our goal in this chapter is to provide a road map for the integration of semantic and ontological knowledge into the process of Web usage mining, and particularly, in its application to Web personalization and recommender systems.

Domain knowledge can be integrated into the Web mining process in many ways. This includes leveraging explicit domain ontologies or implicit domain semantics extracted from the content or the structure of documents or Website. In general, however, this process may involve one or more of three critical activities: domain ontology acquisition, knowledge base construction, and knowledge-enhanced pattern discovery.

**Domain Ontology Acquisition**

The process of acquiring, maintaining and enriching the domain ontologies is referred to as “ontology engineering”. For small Web sites with only static Web pages, it is feasible to construct a domain knowledge base manually or semi-
Integrating Semantic Knowledge with Web Usage Mining for Personalization

manually. In Loh, Wives and de Oliveira (2000), a semi-manual approach is adopted for defining each domain concept as a vector of terms with the help of existing vocabulary and natural language processing tools.

However, manual construction and maintenance of domain ontologies requires a great deal of effort on the part of knowledge engineers, particularly for large-scale Websites or Websites with dynamically generated content. In dynamically generated Websites, page templates are usually populated based on structured queries performed against back-end databases. In such cases, the database schema can be used directly to acquire ontological information. Some Web servers send structured data files (e.g., XML files) to users and let client-side formatting mechanisms (e.g., CSS files) work out the final Web representation on clients. In this case, it is generally possible to infer the schema from the structured data files.

When there is no direct source for acquiring domain ontologies, machine learning and text mining techniques must be employed to extract domain knowledge from the content or hyperlink structure of the Web pages. In Clerkin, Cunningham, and Hayes (2001), a hierarchical clustering algorithm is applied to terms in order to create concept hierarchies. In Stumme, Taouil, Bastide, Pasquier and Lakhal (2000) a Formal Concept Analysis framework is proposed to derive a concept lattice (a variation of association rule algorithm). The approach proposed in Maedche and Staab (2000) learns generalized conceptual relations by applying association rule mining. All these efforts aim to automatically generate machine understandable ontologies for Website domains.

The outcome of this phase is a set of formally defined domain ontologies that precisely represent the Website. A good representation should provide machine understandability, the power of reasoning, and computation efficiency. The choice of ontology representation language has a direct effect on the flexibility of the data mining phase. Common representation approaches are vector-space model (Loh et al., 2000), descriptive logics (such as DAML+OIL) (Giugno & Lukasiewicz, 2002; Horrocks & Sattler, 2001), first order logic (Craven et al., 2000), relational models (Dai & Mobasher, 2002), probabilistic relational models (Getoor, Friedman, Koller & Taskar, 2001), and probabilistic Markov models (Anderson, Domingos & Weld, 2002).

Knowledge Base Construction

The first phase generates the formal representation of concepts and relations among them. The second phase, knowledge base construction, can be viewed as building mappings between concepts or relations on the one hand, and objects on the Web. The goal of this phase is to find the instances of the concepts and relations from the Website’s domain, so that they can be exploited to perform further data mining tasks. Learning algorithms play an important role in this phase.

In Ghani and Fano (2002), a text classifier is learned for each “semantic feature” (somewhat equivalent to the notion of a concept) based on a small manually labeled data set. First, Web pages are extracted from different Websites that belong to a similar domain, and then the semantic features are manually labeled. This small labeled data set is fed into a learning algorithm as the training data to learn the mappings between Web objects and the concept labels. In fact, this approach treats the process of assigning concept labels as filling “missing” data. Craven et al. (2000) adopt a combined approach of statistical text classification and first order text classification in recognizing concept instances. In that study, learning process is based on both page content and linkage information.

Knowledge-Enhanced Web Data Mining

Domain knowledge enables analysts to perform more powerful Web data mining tasks. The ap-
Applications include content mining, information retrieval and extraction, Web usage mining, and personalization. On the other hand, data mining tasks can also help to enhance the process of domain knowledge discovery.

Domain knowledge can improve the accuracy of document clustering and classification and induce more powerful content patterns. For example, in Horrocks (2002), domain ontologies are employed in selecting textual features. The selection is based on lexical analysis tools that map terms into concepts within the ontology. The approach also aggregates concepts by merging the concepts that have low support in the documents. After preprocessing, only necessary concepts are selected for the content clustering step. In McCullum, Rosenfeld, Mitchell and Ng (1998), a concept hierarchy is used to improve the accuracy and the scalability of text classification.

Traditional approaches to content mining and information retrieval treat every document as a set or a bag of terms. Without domain semantics, we would treat “human” and “mankind” as different terms, or, “brake” and “car” as unrelated terms. In Loh et al. (2000), a concept is defined as a group of terms that are semantically relevant, for example, as synonyms. With such concept definitions, concept distribution among documents is analyzed to find interesting concept patterns. For example, one can discover dominant themes in a document collection or in a single document; or find associations among concepts.

Ontologies and domain semantics have been applied extensively in the context of Web information retrieval and extraction. For example, the ARCH system (Parent, Mobasher & Lytinen, 2001) adopts concept hierarchies because they allow users to formulate more expressive and less ambiguous queries when compared to simple keyword-based queries. In ARCH, an initial user query is used to find matching concepts within a portion of concept hierarchy. The concept hierarchy is stored in an aggregated form with each node represented as a term vector. The user can select or unselect nodes in the presented portion of the hierarchy, and relevance feedback techniques are used to modify the initial query based on these nodes.

Similarly, domain-specific search and retrieval applications allow for a more focused and accurate search based on specific relations inherent in the underlying domain knowledge. The CiteSeer system (Bollacker, Lawrence & Giles, 1998) is a Web agent for finding interesting research publications, in which the relation “cited by” is the primary relation discovered among objects (i.e., among published papers). Thus, CiteSeer allows for comparison and retrieval of documents, not only based on similar content, but also based on inter-citation linkage structure among documents.

CiteSeer is an example of an approach for integrating semantic knowledge based on Web structure mining. In general, Web structure mining tasks take as input the hyperlink structure of Web pages (either belonging to a Website or relative to the whole Web), and output the underlying patterns (e.g., page authority values, linkage similarity, Web communities, etc.) that can be inferred from the hypertext co-citations. Another example of such an approach is the PageRank algorithm, which is the backbone of the Google search engine. PageRank uses the in-degree of indexed pages (i.e., number of pages referring to it) in order to rank pages based on quality or authoritativeness. Such algorithms that are based on the analysis of structural attributes can be further enhanced by integrating content semantics (Chakrabarti et al., 1998). Web semantics can also enhance crawling algorithms by combining content or ontology with linkage information (Chakrabarti, van den Berg & Dom, 1999; Maedche, Ehrig, Handschuh, Volz & Stojanovic, 2002).

The use of domain knowledge can also provide tremendous advantages in Web usage mining and personalization. For example, semantic knowledge may help in interpreting, analyzing, and reasoning about usage patterns discovered in the mining
Integrating Semantic Knowledge with Web Usage Mining for Personalization

phase. Furthermore, it can enhance collaborative filtering and personalization systems by providing concept-level recommendations (in contrast to item-based or user-based recommendations). Another advantage is that user demographic data, represented as part of a domain ontology, can be more systematically integrated into collaborative or usage-based recommendation engines. Several studies have considered various approaches to integrate content-based semantic knowledge into traditional collaborative filtering and personalization frameworks (Anderson et al., 2002; Claypool et al., 1999; Melville, Mooney & Nagarajan, 2002; Mobasher, Dai, Luo, Sun & Zhu, 2000b; Pazzani, 1999). Recently, we proposed a formal framework for integrating full domain ontologies with the personalization process based on Web usage mining (Dai & Mobasher, 2002).

WEB USAGE MINING AND PERSONALIZATION

The goal of personalization based on Web usage mining is to recommend a set of objects to the current (active) user, possibly consisting of links, ads, text, products, and so forth, tailored to the user’s perceived preferences as determined by the matching usage patterns. This task is accomplished by matching the active user session (possibly in conjunction with previously stored profiles for that user) with the usage patterns discovered through Web usage mining. We call the usage patterns used in this context aggregate usage profiles since they provide an aggregate representation of the common activities or interests of groups of users. This process is performed by the recommendation engine which is the online component of the personalization system. If the data collection procedures in the system include the capability to track users across visits, then the recommendations can represent a longer term view of user’s potential interests based on the user’s activity history within the site. If, on the other hand, aggregate profiles are derived only from user sessions (single visits) contained in log files, then the recommendations provide a “short-term” view of user’s navigational interests. These recommended objects are added to the last page

Figure 1. General Framework for Web Personalization Based on Web Usage Mining — The Offline Pattern Discovery Component
Integrating Semantic Knowledge with Web Usage Mining for Personalization

in the active session accessed by the user before that page is sent to the browser.

The overall process of Web personalization based on Web usage mining consists of three phases: data preparation and transformation, pattern discovery, and recommendation. Of these, only the latter phase is performed in real-time. The data preparation phase transforms raw Web log files into transaction data that can be processed by data mining tasks. A variety of data mining techniques can be applied to this transaction data in the pattern discovery phase, such as clustering, association rule mining, and sequential pattern discovery. The results of the mining phase are transformed into aggregate usage profiles, suitable for use in the recommendation phase. The recommendation engine considers the active user session in conjunction with the discovered patterns to provide personalized content.

The primary data sources used in Web usage mining are the server log files, which include Web server access logs and application server logs. Additional data sources that are also essential for both data preparation and pattern discovery include the site files (HTML, XML, etc.) and meta-data, operational databases, and domain knowledge. Generally speaking, the data obtained through these sources can be categorized into four groups (see also Cooley, Mobasher, & Srivastava, 1999; Srivastava et al., 2000).

- **Usage data:** The log data collected automatically by the Web and application servers represents the fine-grained navigational behavior of visitors. Depending on the goals of the analysis, this data needs to be transformed and aggregated at different levels of abstraction. In Web usage mining, the most basic level of data abstraction is that of a pageview. Physically, a pageview is an aggregate representation of a collection of Web objects contributing to the display on
a user’s browser resulting from a single user action (such as a clickthrough). These Web objects may include multiple pages (such as in a frame-based site), images, embedded components, or script and database queries that populate portions of the displayed page (in dynamically generated sites). Conceptually, each pageview represents a specific “type” of user activity on the site, e.g., reading a news article, browsing the results of a search query, viewing a product page, adding a product to the shopping cart, and so on. On the other hand, at the user level, the most basic level of behavioral abstraction is that of a server session (or simply a session). A session (also commonly referred to as a “visit”) is a sequence of pageviews by a single user during a single visit. The notion of a session can be further abstracted by selecting a subset of pageviews in the session that are significant or relevant for the analysis tasks at hand. We shall refer to such a semantically meaningful subset of pageviews as a transaction. It is important to note that a transaction does not refer simply to product purchases, but it can include a variety of types of user actions as captured by different pageviews in a session.

• **Content data:** The content data in a site is the collection of objects and relationships that are conveyed to the user. For the most part, this data is comprised of combinations of textual material and images. The data sources used to deliver or generate this data include static HTML/XML pages, image, video, and sound files, dynamically generated page segments from scripts or other applications, and collections of records from the operational database(s). The site content data also includes semantic or structural meta-data embedded within the site or individual pages, such as descriptive keywords, document attributes, semantic tags, or HTTP variables. Finally, the underlying domain ontology for the site is also considered part of content data. The domain ontology may be captured implicitly within the site, or it may exist in some explicit form. The explicit representation of domain ontologies may include conceptual hierarchies over page contents, such as product categories, structural hierarchies represented by the underlying file and directory structure in which the site content is stored, explicit representation of semantic content and relationships via an ontology language such as RDF, or a database schema over the data contained in the operational databases.

• **Structure data:** The structure data represents the designer’s view of the content organization within the site. This organization is captured via the inter-page linkage structure among pages, as reflected through hyperlinks. The structure data also includes the intra-page structure of the content represented in the arrangement of HTML or XML tags within a page. For example, both HTML and XML documents can be represented as tree structures over the space of tags in the page. The structure data for a site is normally captured by an automatically generated “site map” which represents the hyperlink structure of the site. A site mapping tool must have the capability to capture and represent the inter- and intra-pageview relationships. This necessity becomes most evident in a frame-based site where portions of distinct pageviews may represent the same physical page. For dynamically generated pages, the site mapping tools must either incorporate intrinsic knowledge of the underlying applications and scripts, or must have the ability to generate content segments using a sampling of parameters passed to such applications or scripts.

• **User data:** The operational database(s) for the site may include additional user profile information. Such data may include demo-
graphic or other identifying information on registered users, user ratings on various objects such as pages, products, or movies, past purchase or visit histories of users, as well as other explicit or implicit representation of users’ interests. Obviously, capturing such data would require explicit interactions with the users of the site. Some of this data can be captured anonymously, without any identifying user information, so long as there is the ability to distinguish among different users. For example, anonymous information contained in client-side cookies can be considered a part of the user’s profile information, and can be used to identify repeat visitors to a site. Many personalization applications require the storage of prior user profile information. For example, collaborative filtering applications, generally, store prior ratings of objects by users, though, such information can be obtained anonymously, as well.

For a detailed discussion of preprocessing issues related to Web usage mining see Cooley et al. (1999). Usage preprocessing results in a set of \( n \) pageviews, \( P = \{ p_1, p_2, \ldots, p_n \} \), and a set of \( m \) user transactions, \( T = \{ t_1, t_2, \ldots, t_m \} \), where each \( t \) in \( T \) is a subset of \( P \). Pageviews are semantically meaningful entities to which mining tasks are applied (such as pages or products). Conceptually, we view each transaction \( t \) as an \( l \)-length sequence of ordered pairs:

\[
t = \left( \left( p_{t_1}, w(p_{t_1}) \right), \left( p_{t_2}, w(p_{t_2}) \right), \ldots, \left( p_{t_l}, w(p_{t_l}) \right) \right)
\]

where each \( p_{t_j} = p_j \) for some \( j \) in \( \{1, \ldots, n\} \), and \( w(p_j) \) is the weight associated with pageview \( p_j \) in transaction \( t \), representing its significance (usually, but not exclusively, based on time duration).

For many data mining tasks, such as clustering and association rule discovery, as well as collaborative filtering based on the \( k \)-NN technique, we can represent each user transaction as a vector over the \( n \)-dimensional space of pageviews. Given the transaction \( t \) above, the transaction vector \( \tilde{t} \) is given by: \( \tilde{t} = \{ w_{p_1}, w_{p_2}, \ldots, w_{p_n} \} \), where each \( w_{p_j} = w(p_j) \), for some \( j \) in \( \{1, \ldots, n\} \), if \( p_j \) appears in the transaction \( t \), and \( w_{p_j} = 0 \), otherwise. Thus, conceptually, the set of all user transactions can be viewed as an \( m \times n \) transaction-pageview matrix, denoted by \( TP \).

Given a set of transactions as described above, a variety of unsupervised knowledge discovery techniques can be applied to obtain patterns. These techniques such as clustering of transactions (or sessions) can lead to the discovery of important user or visitor segments. Other techniques such as item (e.g., pageview) clustering and association or sequential pattern discovery can be used to find important relationships among items based on the navigational patterns of users in the site. In each case, the discovered patterns can be used in conjunction with the active user session to provide personalized content. This task is performed by a recommendation engine.

**Requirements for Semantic Web Usage Mining**

In this section, we present and discuss the essential requirements in the integration of domain knowledge with Web usage data for pattern discovery. Our focus is on the critical tasks that particularly play an important role when the discovered patterns are to be used for Web personalization. As a concrete example, in the last part of this section we discuss an approach for integrating semantic features extracted from the content of Web sites with Web usage data, and how this integrated data can be used in conjunction with clustering to perform personalization. In the next section, we go beyond keyword-based semantics and present a more formal framework for integrating full ontologies with the Web usage mining and personalization processes.
Representing Domain Knowledge as Content Features

One direct source of semantic knowledge that can be integrated into mining and personalization processes is the textual content of Web site pages. The semantics of a Web site are, in part, represented by the content features associated with items or objects on the Web site. These features include keywords, phrases, category names, or other textual content embedded as meta-information. Content preprocessing involves the extraction of relevant features from text and meta-data.

During the preprocessing, usually different weights are associated with features. For features extracted from meta-data, feature weights are usually provided as part of the domain knowledge specified by the analyst. Such weights may reflect the relative importance of certain concepts. For features extracted from text, weights can normally be derived automatically, for example as a function of the term frequency and inverse document frequency (tf.idf) which is commonly used in information retrieval.

Further preprocessing on content features can be performed by applying text mining techniques. This would provide the ability to filter the input to, or the output from, other mining algorithms. For example, classification of content features based on a concept hierarchy can be used to limit the discovered patterns from Web usage mining to those containing pageviews about a certain subject or class of products. Similarly, performing learning algorithms such as, clustering, formal concept analysis, or association rule mining on the feature space can lead to composite features representing concept categories or hierarchies (Clerkin, Cunningham, & Hayes, 2001; Stumme et al., 2000).

The integration of content features with usage-based personalization is desirable when we are dealing with sites where text descriptions are dominant and other structural relationships in the data are not easy to obtain, e.g., news sites or online help systems, etc. This approach, however, is incapable of capturing more complex relations among objects at a deeper semantic level based on the inherent properties associated with these objects. To be able to recommend different types of complex objects using their underlying properties and attributes, the system must be able to rely on the characterization of user segments and objects, not just based on keywords, but at a deeper semantic level using the domain ontologies for the objects. We will discuss some examples of how integrated content features and usage data can be used for personalization later in this Section.

Representing Domain Knowledge as Structured Data

In Web usage mining, we are interested in the semantics underlying a Web transaction or a user profile which is usually composed of a group of pageview names and query strings (extracted from Web server logs). Such features, in isolation, do not convey the semantics associated with the underlying application. Thus, it is important to create a mapping between these features and the objects, concepts, or events they represent.

Many e-commerce sites generate Web pages by querying operational databases or semi-structured data (e.g., XML and DTDs), from which semantic information can be easily derived. For Web sites in which such structured data cannot be easily acquired, we can adopt machine learning techniques to extract semantic information. Furthermore, the domain knowledge acquired should be machine understandable in order to allow for further processing or reasoning. Therefore, the extracted knowledge should be represented in some standard knowledge representation language.

DAML+OIL (Horrocks & Sattler, 2001) is an example of an ontology language that combines
the Web standards from XML and RDF, with the reasoning capabilities from a description logic \textit{SHIP(DL)}. The combinations of relational models and probabilistic models is another common approach to enhance Web personalization with domain knowledge and reasoning mechanism. Several approaches to personalization have used Relational Models such as Relational Markov Model (Anderson et al., 2002). Both of these approaches provide the ability to represent knowledge at different levels of abstraction, and the ability to reason about concepts, including about such relations as subsumption and membership.

In Dai & Mobasher (2002), we adopted the syntax and semantics of another ontology representation framework, \textit{SHOQ(D)}, to represent domain ontologies. In \textit{SHOQ(D)}, the notion of \textit{concrete datatype} is used to specify literal values and \textit{individuals} which represent real objects in the domain ontology. Moreover, \textit{concepts} can be viewed as sets of individuals, and \textit{roles} are binary relations between a pair of concepts or between concepts and data types. The detailed formal definitions for concepts and roles are given in Horrocks & Sattler (2001) and Giugno & Lukasiewicz (2002). Because our current work does not focus on reasoning tasks such as deciding subsumption and membership, we do not focus our discussion on these operations. The reasoning apparatus in \textit{SHOQ(D)} can be used to provide more intelligent data mining services.

Building “Mappings” Between Usage-Level and Domain-Level Instances

During usage data preprocessing or post processing, we may want to assign domain semantics to user navigational patterns by mapping the pageview names or URLs (or queries) to the instances in the knowledge base. To be more specific, instead of describing a user’s navigational path as: “\(a_1, a_2, \ldots, a_n\)” (where \(a_i\) is a URL pointing to a Web resource), we need to represent it using the instances from the knowledge base, such as: “\(\text{movie(name=Matrix)}, \text{movie(name=Spiderman)}, \ldots, \text{movie(name=Xman)}\).” With the help of a pre-acquired concept hierarchy, we may, for example, be able to infer that the current user’s interest is in the category of “Action\&Sci-Fi.” We refer to this “semantic” form of usage data as “Semantic User Profiles.” These profiles, in turn, can be used for semantic pattern discovery and online recommendations. In the context of personalization applications, domain-level (semantic) instances may also need to be mapped back to Web resources or pages. For example, a recommendation engine using semantic user profiles may result in recommendations in the form of a movie genre. This concept must be mapped back into specific pages, URLs, or sections of the site relating to this genre before recommendations can be relayed to the user.

Using Content and Structural Characteristics

Classification algorithms utilizing content and structural features from pages are well-suited for creating mappings from usage data to domain-level instances. For example, in Craven et al. (2000) and Ghani & Fano (2002) classifiers are trained that exploit content or structural features (such as terms, linkage information, and term proximity) of the pageviews. From the pageview names or URLs we can obtain the corresponding Web content such as meta-data or keywords. With help from text classification algorithms, it is possible to efficiently map from keywords to attribute instances (Ghani & Fano, 2002).

Another good heuristics used in creating semantic mappings is based on the anchor text associated with hyperlinks. If we can build the complete user navigational path, we would be able to acquire the anchor text for each URL or pageview name. We can include the anchor text as part of the content features extracted from the body of documents or in isolation. However,
whereas the text features in a document represent the semantics of the document, itself, the anchor text represents the semantics of the document to which the associated hyperlink points.

Using Query Strings

So far, we have overlooked the enormous amount of information stored in databases or semi-structured documents associated with a site. Large information servers often serve content integrated from multiple underlying servers and databases (Berendt, & Spiliopoulou, 2000). The dynamically generated pages on such servers are based on queries with multiple parameters attached to the URL corresponding to the underlying scripts or applications. Using the Web server query string recorded in the server log files it is possible to reconstruct the response pages. For example, the following are query strings from a hypothetical online bookseller Web site:

http://www.xyz.com/app.cgi?action=viewitem&item=1234567&category=1234

http://www.xyz.com/app.cgi?action=search&searchtype=title&searchstring=web+mining

http://www.xyz.com/app.cgi?action=order&item=1234567&category=1234&couponid=3456

If the background database or semi-structured documents are available, then we can access the content of the instances in the response pages via the name-value pairs from the query strings. This enriches our knowledge base of user interest. In the above bookseller Web site example, if we were able to access background database, we would be able to get the content of item “1234567” in category “1234”. In this case, we could have the book name, price, author information of this item. We could recommend other books in the same content category or written by the same author. More generally, in well-designed sites, there is usually an explicitly available semantic mapping between query parameters and objects (such as products and categories), which would obviate the need to reconstruct the content of dynamic pages.

Levels of Abstraction

Capturing semantic knowledge at different levels of abstraction provides more flexibility both in the mining phase and in the recommendation phase. For example, focusing on higher-level concepts in a concept hierarchy would allow certain patterns to emerge which otherwise may be missed due to low support. On the other hand, the ability to drill-down into the discovered patterns based on finer-grained subconcepts would provide the ability to give more focused and useful recommendations.

Domain knowledge with attributes and relations requires the management of a great deal more data than is necessary in traditional approaches to Web usage mining. Thus, it becomes essential to prune unnecessary attributes or relations. For example, it may be possible to examine the number of distinct values of each attribute and generalize the attributes if there is a concept hierarchy over the attribute values. In Han & Fu (1995) a multiple-level association rule mining algorithm is proposed that utilizes concept hierarchies. For example, the usage data in our hypothetical movie site may not provide enough support for an association rule: “Spiderman, Xmen → Xmen2”, but mining at a higher level may result in obtaining a rule: “Sci-Fi&Action, Xmen → Xmen2”. In Anderson et al. (2002) relational Markov models are built by performing shrinkage (McCallum et al., 1998) between the estimates of parameters at all levels of abstractions relative to a concept hierarchy. If a pre-specified concept hierarchy does not exist, it is possible to automatically create such hierarchies through a variety of machine learning techniques, such as hierarchical agglomerative clustering (Stumme et al., 2000).
Integration of Semantics at Different Stages of Knowledge Discovery

The semantic information stored in the knowledge base can be leveraged at various steps in the knowledge discovery process, namely in the preprocessing phase, in the pattern discovery phase, or during the post-processing of the discovered patterns.

Preprocessing Phase

The main task of data preprocessing is to prune noisy and irrelevant data, and to reduce data volume for the pattern discovery phase. In Mobasher, Dai, Luo, & Nakagawa (2002), it was shown that applying appropriate data preprocessing techniques on usage data could improve the effectiveness of Web personalization. The concept level mappings from the pageview-level data to concepts can also be performed in this phase. This results in a transformed transaction data to which various data mining algorithms can be applied. Specifically, the transaction vector $t$ given previously can be transformed into a vector $t' = \{w'_o, w'_o, \ldots, w'_o\}$, where each $o_j$ is a semantic object appearing in one of the pageviews contained in the transaction, and is a weight associated with that object in the transaction. These semantic objects may be concepts appearing in the concept hierarchy or finer-grained objects representing instances of these concepts.

Pattern Discovery Phase

Successful utilization of domain knowledge in this phase requires extending basic data mining algorithms to deal with relational data to concept hierarchies. As an example, consider a distance-based data mining technique such as clustering. The clustering of flat single-relation data (such as Web user transactions) involves the computation of similarities or distance among transaction vectors. In such cases, normally simple vector-based operations are used. However, in the presence of integrated domain knowledge represented as concept hierarchies or ontologies, the clustering algorithms will have to perform much more complex similarity computations across dimensions and attributes. For example, even if the two user transactions have no pageviews in common, they may still be considered similar provided that the items occurring in both transactions are themselves “similar” based on some of their attributes or properties. The integration of domain knowledge will generate “semantic” usage patterns, introducing great flexibility as well as challenges. The flexibility lies in the pattern discovery being independent of item identities. The challenge is in the development of scalable and efficient algorithms to perform the underlying computational tasks such as similarity computations. We discuss this issue further below.

Post-Processing Phase

Exploiting domain knowledge in this phase can be used to further explain usage patterns or to filter out irrelevant patterns. One possibility is to first perform traditional usage mining tasks on the item-level usage data obtained in the preprocessing phase, and then use domain knowledge to interpret or transform the item level user profiles into “domain-level usage profiles” (Mobasher & Dai, 2002) involving concepts and relations in the ontology. The advantage of this approach is that we can avoid the scalability issues that can be endemic in the pattern discovery phase. The disadvantage is that some important structural relationships may not be used during the mining phase resulting in lower quality patterns.

Aggregation Methods for Complex Objects

To characterize patterns discovered through data mining techniques, it is usually necessary to de-
derive aggregate representation of the patterns. An example of this situation is clustering applications. In the context of Web user transactions, clustering may result in a group of sessions or visitors that are considered similar because of their common navigational patterns. The vector representation of these transactions facilitates the aggregation tasks: the centroid (mean vector) of the transaction cluster acts as a representative of all of the transactions in that cluster. However, in the case of semantically enhanced transactions, the aggregation may have to be performed independently for each of the attributes associated with the objects contained in the cluster.

For example, clustering may result in a group of users who have all visited pages related to several movies. To be able to characterize this group of users at a deeper semantic level, it would be necessary to create an aggregate representation of the collection of movies in which they are interested. This task would require aggregation along each dimension corresponding to the attributes of “movie” instances, such as “genre”, “actors”, “directors”, etc. Since each of these attributes require a different type of aggregation function depending on the data type and the domain, it may be necessary to associate various aggregation functions with the specification of the domain ontology, itself. In the next section we present one approach for solving this problem.

**Measuring Semantic Similarities**

Measuring similarities (alternatively, distances) among objects is a central task in many data mining algorithms. In the context of Web usage mining this may involve computing similarity measures among pageviews, among user transactions, or among users. This also becomes a critical task in personalization: a current user’s profile must be matched with similar aggregate profiles representing the discovered user patterns or segments. As in the case of the aggregation problem discussed above, when dealing with semantically enhanced transactions, measuring similarities poses additional challenges. This is because the similarity of two transactions depends on the similarities of the semantic objects contained within the transactions.

Let us again consider the static vector model for representing a Web transaction $t$ (or a user profile): $t = \langle w'_1, w'_2, \ldots, w'_p \rangle$. Computing similarity between two such vectors is straightforward and can be performed using measures such as cosine similarity, Euclidean distance, Pearson correlation (e.g., in case the weights represent user ratings).

When such vectors are transformed according to the underlying semantics, however, the computation of similarities will involve the computation of semantic similarities among the concepts or objects, possibly using different domain-specific similarity measures. Let $A$ and $B$ be two transformed transactions, each represented as a set of semantic objects in a site:

$A = \{a_1, a_2, \ldots, a_m\}$ and $B = \{b_1, b_2, \ldots, b_l\}$.

The computation of vector similarity between $A$ and $B$, $Sim(A,B)$, is dependent on the semantic similarities among the component objects, $SemSim(a_i,b_j)$. For instance, one approach might be to compute the weighted sum or average of the similarities among object pairs, such as in:

$$Sim(A,B) = \sum_{a_i \in A} \sum_{b_j \in B} \frac{SemSim(a_i,b_j)}{|A||B|}$$

In general, computing the semantic similarity, $SemSim(a,b)$, is domain dependent and requires knowledge of the underlying structure of among objects. If both objects can be represented using the same vector model (e.g., pages or documents represented as bags of words), we can compute their similarity using standard vector operations. On the other hand, if their representation includes attributes and relations specified in the domain ontology, we need to first make sure that the ob-
jects can be classified under a common ontological schema and then measure similarities along the different dimensions corresponding to each attribute. The notion of semantic matching among objects and classes has been a subject of considerable study recently (Rodriguez & Egenhofer, 2003; Palopoli, Sacca, Terracina, & Ursino, 2003; Ganesan, Garcia-Molina, & Widom, 2003).

For example, such an approach was used in Jin & Mobasher (2003) in the context of collaborative filtering with movies. In this work, association analysis was first performed on the “genre” attribute to define a genre hierarchy. Furthermore, the “year” attribute was discretized into intervals, while other attributes, such as “cast”, were treated as a bag of words. These preprocessing steps allowed for the definition of appropriate similarity measures for each attribute. Finally, the semantic similarity between two movies, $i$ and $j$, was defined as a linear combination of attribute-level similarities:

$$
SemSim(i, j) = \alpha_1 \ast CastSim(i, j) + \alpha_2 \ast DirectorSim(i, j) + \alpha_3 \ast GenreSim(i, j) + \ldots
$$

where, $\alpha_i$ are predefined weights for the corresponding attributes.

**Example: Using Content Features for Semantic Web Usage Mining**

As an example of integrating semantic knowledge with the Web usage mining process, let us consider the especial case of using textual features from the content of Web pages to represent the underlying semantics for the site. As noted earlier, each pageview $p$ can be represented as a $k$-dimensional feature vector, where $k$ is the total number of extracted features (words or concepts) from the site in a global dictionary. This vector can be given by:

$$p = \{fw(p, f_1), fw(p, f_2), \ldots, fw(p, f_k)\}$$

where $fw(p, f_j)$ is the weight of the $j$th feature in pageview $p$, for $1 \leq j \leq k$. For the whole collection of pageviews in the site, we then have an $n \times k$ pageview-feature matrix $PF = \{p_1, p_2, \ldots, p_n\}$.

There are now at least two basic choices as to when content features can be integrated into the usage-based personalization process: pre-mining integration or post-mining integration.

The pre-mining integration involves the transformation of user transactions, as described earlier, into “content-enhanced” transactions containing the semantic features of the pageviews. While, in practice, there are several ways to accomplish this transformation, the most direct approach involves mapping each pageview in a transaction to one or more content features. The range of this mapping can be the full feature space, or feature sets (composite features) which in turn may represent concepts and concept categories. Conceptually, the transformation can be viewed as the multiplication of the transaction-pageview matrix $TP$, defined earlier, with the pageview-feature matrix $PF$. The result is a new matrix $TF = \{t_1, t_2, \ldots, t_n\}$, where each $t_i$ is a $k$-dimensional vector over the feature space. Thus, a user transaction can be represented as a content feature vector, reflecting that user’s interests in particular concepts or topics.

Various data mining tasks can now be performed on the content-enhanced transaction data. For instance, if we apply association rule mining to such data, then we can get a group of association rules on content features. As an example, consider a site containing information about movies. This site may contain pages related to the movies themselves, actors appearing in the movies, directors, and genres. Association rule mining process could generate a rule such as: (“British”, “Romance”, “Comedy” $\Rightarrow$ “Hugh Grant”), suggesting that users who are interested in British romantic comedies may also like the actor Hugh Grant (with a certain degree of confidence). During the online recommendation phase, the user’s active session (which is also transformed into a feature vector) is compared with the discovered
rules. Before recommendations are made, the matching patterns must be mapped back into Web pages or Web objects. In the above example, if the active session matches the left hand side of the association rule, the recommendation engine could recommend other Web pages that contain the feature “Hugh Grant”.

The post-mining integration of semantic features involves combining the results of mining (performed independently on usage and content data) during the online recommendation phase. An example of this approach was presented in Mobasher et al. (2000b), where clustering algorithms were applied to both the transaction matrix $TP$ and the transpose of the feature matrix $PF$. Since both matrices have pageviews as dimensions, the centroids of the resulting clusters in both cases can be represented as sets of pageview-weight pairs where the weights signify the frequency of the pageview occurrence in the corresponding cluster. We call the patterns generated from content data “content profiles”, while the patterns derived from usage data are called “usage profiles”. Though they share the same representation, they have different semantics: usage profiles represent a set of transactions with similar navigational behavior, while content profiles contain groups of Web pages with (partly) similar content.

Specifically, given a transaction cluster (respectively, a feature cluster) $cl$, we can construct the usage (respectively, content) profile $pr_{ci}$ as a set of pageview-weight pairs by computing the centroid of $cl$:

$$pr_{ci} = \{(p, weight(p, pr_{ci}))\mid weight(p, pr_{ci}) \geq \mu\}$$

where:

- $w(p,s)$ is the weight of page $p$ in transaction (respectively, feature) vector $s$ in the cluster $cl$; and
- the threshold $\mu$ is used to focus only on those pages in the cluster that appear in a sufficient number of vectors in that cluster.

Each such profile, in turn, can be represented as a vector in the original $n$-dimensional space of pageviews. This aggregate representation can be used directly in the recommendation phase: given a new user, $u$ who has accessed a set of pages, $P_u$, so far, we can measure the similarity of $P_u$ to the discovered profiles, and recommend to the user those pages in matching profiles which have not yet been accessed by the user. Note that this approach does not distinguish between recommendations emanating from the matching content and usage profiles. Also note that there are many other ways of combining usage profiles and content profiles during the online recommendation phase. For example, we can use content profiles as the last resort in the situation where usage profiles can not provide sufficient number of recommendations.

**A FRAMEWORK FOR ONTOLOGY-BASED PERSONALIZATION**

At a conceptual level, there may be many different kinds of objects within a given site that are accessible to users. At the physical level, these objects may be represented by one or more Web pages. For example, our hypothetical movie site may contain pages related to the movies, actors, directors, and studios. Conceptually, each of these entities represents a different type of semantic object. During a visit to this site, a user may implicitly access several of these objects together during a session by navigating to various pages containing them. In contrast to content features, ontological representation of domain knowledge contained in the site makes it possible to have a...
uniform architecture to model such objects, their properties, and their relationships. Furthermore, such a representation would allow for a more natural mapping between the relational schema for the backend databases driving Web applications and the navigational behavior of users.

In this section we will present a general framework for utilizing domain ontologies in Web usage mining and personalization. Figure 3 lays out a general process for such an integrated approach. In keeping with our earlier discussion, it is composed of three main phases: preprocessing, pattern discovery and online recommendation. Each of these phases must take into account the object properties and their relationships as specified in a domain ontology.

We assume that the site ontology is already available (either specified manually, or extracted automatically using ontology learning techniques). The goal of the preprocessing phase is to transform users’ navigational transactions into “semantic transactions” by mapping accessed pages and resource to concepts and objects of the specified ontology. The goal of the pattern discovery phase is to create aggregate representation of groups of semantic objects that are implicitly accessed by similar users, thus providing a semantic characterization of user segments with common behavior or interests. Finally, in the recommendation phase, the discovered semantic patterns are utilized in conjunction with an ongoing record of a current user’s activity (including, possibly the user’s stored profile) to recommend new resources, pages, or objects to that user.

Knowledge Representation

General ontology representation languages such as DAML+OIL (Horrocks, 2002) provide formal syntax and semantics for representing and reasoning with various elements of an ontology. These elements include “individuals” (or objects), “concepts” (which represent sets of individuals), and “roles” (which specify object properties). In DAML+OIL, the notion of a concept is quite general and may encompass a heterogeneous set of objects with different properties (roles) and structures. We, on the other hand, are mainly

Figure 3. A General Framework for Personalization Based on Domain Ontologies
interested in the aggregate representations for groups of objects that have a homogenous concept structure (i.e., have similar properties and data types). For example, we may be interested in a group of movie objects, each of which has specific values for properties such as “year”, “genre”, and “actors.” We call such a group of objects a class. Thus, in our framework, the notion of a class represents a restriction of the notion of a concept in DAML+OIL. It should be noted, however, that the users of a Web site, in general, access a variety of objects belonging to different classes. Thus, this homogeneity assumption would imply that semantic objects within user transactions must first be classified into homogenous classes as a preprocessing step.

More specifically, we define a class \( C \) as a set of objects together with a set of attributes. These attributes together define the internal properties of the objects in \( C \) or relationships with other concepts that involve the objects in \( C \). Thus attributes of a class correspond to a subset of the set of roles in the domain ontology. We denote the domain of values of an attribute \( a \) as \( D_a \). Furthermore, because we are specifically interested in aggregating objects at the attribute level, we extend the notion of a role to include a domain-specific combination function and an ordering relation.

More formally, a class \( C \) is characterized by a finite set of attributes \( AC \), where each attribute \( a \) in \( AC \) is defined as follows.

**Definition:** Let \( C \) be a class in the domain ontology. An attribute \( a \) \( AC \) is a 4-tuple, where \( a = (T_a, D_a, \prec_a, \psi_a) \)

- \( T_a \) is the type for the values for the attribute \( a \).
- \( D_a \) is the domain of the values for \( a \);
- \( \prec_a \) is an ordering relation among the values in \( D_a \); and
- \( \psi_a \) is a combination function for the attribute \( a \).

The “type” of an attribute in the above definition may be a concrete datatype (such as “string” or “integer”) or it may be a set of objects (individuals) belonging to another class.

In the context of data mining, comparing and aggregating values are essential tasks. Therefore, ordering relations among values are necessary properties for attributes. We associate an ordering relation \( \prec_a \) with elements in \( D_a \) for each attribute \( a \). The ordering relation \( \prec_a \) can be null (if no ordering is specified in the domain of values), or it can define a partial or a total order among the domain values. For standard types such as values from a continuous range, we assume the usual ordering. In cases when an attribute \( a \) represents a concept hierarchy, the domain values of \( a \) are a set of labels, and \( \prec_a \) is a partial order representing the “is-a” relation.

Furthermore, we associate a data mining operator, called the combination function, \( \psi_a \), with each attribute \( a \). This combination function defines an aggregation operation among the corresponding attribute values of a set of objects belonging to the same class. This function is essentially a generalization of the “mean” or “average” function applied to corresponding dimension values of a set of vectors when computing the centroid vector. In this context, we assume that the combination function is specified as part of the domain ontology for each attribute of a class. An interesting extension would be to automatically learn the combination function for each attribute based on a set of positive and negative examples.

Classes in the ontology define the structural and semantic properties of objects in the domain which are “instances” of that class. Specifically, each object \( o \) in the domain is also characterized by a set of attributes \( A_o \) corresponding to the attributes of a class in the ontology. In order to more precisely define the notion of an object as an instance of a class, we first define the notion of an instance of an attribute.
Definition: Given an attribute \( a = \langle T_a, D_a, \prec_a, \psi_a \rangle \) and an attribute \( b = \langle T_b, D_b, \prec_b, \psi_b \rangle \), \( b \) is an instance of \( a \), if \( D_b \subseteq D_a \), \( T_b = T_a \), \( \psi_b = \psi_a \), and \( \prec_b \) is a restriction of \( \prec_a \) to \( D_b \). The attribute \( b \) is a null instance of \( a \), if \( D_b \) is empty.

Definition: Given a class \( C \) with attribute set \( A_C = \{ a_1^C, a_2^C, \ldots, a_n^C \} \), we say that an object \( o \) is an instance of \( C \), if \( o \) has attributes \( A_o = \{ a_1^o, a_2^o, \ldots, a_n^o \} \) such that each is a, possibly null, instance of \( a_i^C \).

Based on the definitions of attribute and object instances, we can now provide a more formal representation of the combination function \( \psi_a \). Let \( C \) be a class and \( \{ o_1, o_2, \ldots, o_m \} \) a set of object instances of \( C \). Let \( a \in A_C \) be an attribute of class \( C \). The combination function \( \psi_a \) can be represented by:

\[
\psi_a \left( \left\{ \langle a_1^o, w_1 \rangle, \langle a_2^o, w_2 \rangle, \ldots, \langle a_n^o, w_n \rangle \right\} \right) = \left\{ \langle a_{agg}, w_{agg} \rangle \right\}
\]

where each \( o_i \) belonging to object \( o_i \) is an instance of the attribute \( a_i \), and each \( w_i \) is a weight associated with that attribute instance (representing the significance of that attribute relative to the other instances). Furthermore, \( a_{agg} \) is a pseudo instance of \( a \) meaning that it is an instance of \( a \) which does not belong to a real object in the underlying domain. The weight \( w_{agg} \) of \( a_{agg} \) is a function of \( w_1, w_2, \ldots, w_n \).

Given a set of object instances, of a class \( C \), a domain-level aggregate profile for these instances is obtained by applying the combination function for each attribute in \( C \) to all of the corresponding attribute instances across all objects \( o_1, o_2, \ldots, o_m \).

**Ontology Preprocessing**

The ontology preprocessing phase takes as input domain information (such as database schema and metadata, if any) as well as Web pages, and generates the site ontology. For simple Web sites, ontologies can be easily designed manually or derived semi-automatically from the site content. However, it is more desirable to have automatic ontology acquisition methods for a large Web site, especially in Web sites with dynamically generated Web pages. E-commerce Web sites, for instance, usually have well-structured Web content, including predefined metadata or database schema. Therefore it is easier to build automatic ontology extraction mechanisms that are site-specific.

There have been a number of efforts dealing with the ontology learning problem (Clerkin et al., 2001; Craven et al., 2000; Maedche & Staab, 2000). A wide range of information, such as thesauri, content features, and database schema can help to identify ontologies. Many of these approaches have focused on extracting ontological information from the Web, in general. In Berendt et al. (2002) the notion of “Semantic Web Mining” was introduced, including a framework for the extraction of a concept hierarchy and the application of data mining techniques to find frequently occurring combinations of concepts.

**An Example**

As an example, let us revisit our hypothetical movie Web site. The Web site includes collections of pages containing information about movies, actors, directors, etc. A collection of pages describing a specific movie might include information such as the movie title, genre, starring actors, director, etc. An actor or director’s information may include name, filmography (a set of movies), gender, nationality, etc. The portion of domain ontology for this site, as described, contains the classes Movie, Actor and Director (Figure 4). The collection of Web pages in the site represents a group of embedded objects that are the instances of these classes.

In our example, the class Movie has attributes such as Year, Actor (representing the relation
“acted by”), Genre, and Director. The Actor and Director attributes have values that are other objects in the ontology, specifically, object instances of classes Actor and Director, respectively. The attribute Year is an example of an attribute whose datatype is positive integers with the usual ordering. The attribute Genre has a concrete datatype whose domain values in $D_{genre}$ are a set of labels (e.g., “Romance” and “Comedy”). The ordering relation $\prec_{genre}$ defines a partial order based on the “is-a” relation among subsets of these labels (resulting in a concept hierarchy of Genres, a portion of which is shown in Figure 4).

Figure 5 shows a Movie instance “About a Boy” and its related attributes and relations extracted from a Web page. The schema of the class Movie is shown at the bottom left portion of the figure. Here we treat the classes Genre and Year as attributes of the class Movie. The instances of the ontology are shown at the bottom right of the figure. The Genre attribute contains a partial order among labels representing a concept hierarchy of movie genres. We use a restriction of this partial order to represent the genre to which the Movie instance belongs. The diagram also shows a keyword bag containing the important keywords in that page.

An attribute $a$ of an object $o$ has a domain $D_a$. In cases when the attribute has unique a value for an object, $D_a$ is a singleton. For example, consider an object instance of class Movie, “About a Boy” (see Figure 5). The attribute Actor contains two objects “H. Grant” and “T. Collette” that are instances of the class Actor (for the sake of presentation we use the actors’ names to stand for the object instances of Actor). Therefore, $D_{Actor} = \{“H. Grant”, “T. Collette”\}$. Also, a real object may have values for only some of the attributes. In this case the other attributes have empty domains. For instance, the attribute Director in the example has an empty domain and is thus not depicted in the figure. We may, optionally, associate a weight with each value in the attribute domain $D_a$ (usually in the range $[0,1]$). This may be useful in capturing the relative importance of each attribute value.

For example, in a given movie the main actors should have higher weights than other actors in the cast. In our example, the object “H. Grant” has weight 0.6 and the object “Toni Collette” has weight 0.4. Unless otherwise specified, we assume that the weight associated with each attribute value is 1. In the object $o$ shown in Figure 5, the domain for the attribute Genre is the set of labels {Genre-All, Action, Romance, Comedy, Romantic Comedy, Black Comedy, Kids & Family}. The ordering relation $\prec_{genre}$ is a restriction of $\prec_{genre}$.
to the subset \{Genre-All, Comedy, Romantic Comedy, Kids & Family\}.

**Pattern Discovery**

As depicted in Figure 3 domain ontologies can be incorporated into usage preprocessing to generate semantic user transactions, or they can be integrated into pattern discovery phase to generate semantic usage patterns. In the following example, we will focus on the latter approach.

Given a discovered usage profile (for example, a set of pageview-weight pairs obtained by clustering user transactions), we can transform it into a domain-level aggregate representation of the underlying objects (Dai & Mobasher, 2002). To distinguish between the representations we call the original discovered pattern an “item-level” usage profile, and we call the new profile based on the domain ontology a “domain-level” aggregate profile. The item-level profile is first represented as a weighted set of objects: $pr = \{(o_1, w_1), (o_2, w_2), ..., (o_n, w_n)\}$ in which each $o_i$ is an object in the underlying domain ontology and $w_i$ represents $o_i$‘s significance in the profile $pr$. Here we assume that, either using manual rules, or through supervised learning methods, we can extract various object instances represented by the pages in the original page- or item-level usage profile. The transformed profile represents a set of objects accessed together frequently by a group of users (as determined through Web usage mining). Objects, in the usage profile, that belong to the same class are combined to form an aggregated pseudo object belonging to that class. An important benefit of aggregation is that the pattern volume is significantly reduced, thus relieving the computation burden for the recommendation engine. Our goal is to create an aggregate representation of this weighted set of objects to characterize the common interests of
the user segment captured by the usage profile at the domain level.

Given the representation of a profile \( pr \) as a weighted set of objects, the objects in \( pr \) may be instances of different classes \( C_1, C_2, \ldots, C_k \) in the ontology. The process of creating a domain-level aggregate profile begins by partitioning \( pr \) into collections of objects with each collection containing all objects that are instances of a specified class (in other words, the process of classifying the object instances in \( pr \)). Let \( G_i \) denote the elements of \( pr \) that are instances of the class \( C_i \).

Having partitioned \( pr \) into \( k \) groups of homogeneous objects, \( G_1, \ldots, G_k \), the problem is reduced to creating aggregate representation of each partition \( G_i \). This task is accomplished with the help of the combination functions for each of the attributes of \( C_i \), some of whose object instances are contained in \( G_i \). Once the representatives for every partition of objects are created, we assign a significance weight to each representative to mark the importance of this group of objects in the profile. In our current approach the significance weight for each representative is computed as the weighted sum of all the object weights in the partition. However, significance weight can be computed using other numeric aggregation functions.

Examples Continued: Generating Domain-Level Aggregate Profiles

To illustrate the semantic aggregation process, let us return to our movie site example. The aggregation process requires that a “combination function” be defined for each attribute of an object in the domain ontology. Figure 6 and 8 show an example of such process. Each movie object has attribute “Name”, “Actor”, “Genre” and “Year”. For the attribute Name, we are interested in all the movie names appearing in the instances. Thus we can define \( \psi_{\text{Name}} \) to be the union operation performed on all the singleton Name attributes of all movie objects. On the other hand, the attribute Actor contains a weighted set of objects belonging to class Actor. In fact, it represents the relation “Starring” between the actor objects and the movie object. In such cases we can use a vector-based weighted mean operation as the combination function. For example, we will determine the aggregate weight of an actor object \( o \) by:

\[
w_o = \left( \sum_i w_i \cdot w_n \right) / \sum_i w_i.
\]

Applying \( \psi_{\text{Actor}} \) in our example will result in the aggregate actor object \( \{ (S, 0.58), (T, 0.27), (U, 0.09) \} \). As for the attribute Year, the combination function may create a range of all the Year values appearing in the objects. Another possible solution is to discretize the full Year range into decades and find the most common decades that are in the domains of the attribute. In our example, using the range option, this may result in an aggregate instance [1999, 2002] for the Year attribute.

The attribute Genre of Movie contains a partial order representing a concept hierarchy among different Genre values. The combination function, in this case, can perform tree (or graph) matching to extract the common parts of the conceptual hierarchies among all instances. Extracting the common nodes from this hierarchy may also depend on the weights associated with the original objects leading to different weights on the graph edges. For example, given that the higher weight Movies 1 and 2 have “Romance” in common, this node may be selected for the aggregate instance, even though it is not present in Movie 3. However, the weight of “Romance” may be less than that of “Comedy” which is present in all three movies.

Figure 6 shows the item-level usage profile and its representation as a weighted set of objects, as well as the resulting domain-level aggregate profile. Note that the original item-level profile gives us little information about the reasons why these objects were commonly accessed together. However, after we characterize this profile at the domain-level, we find some interesting patterns:
they all belong to Genre “Comedy” (and to a lesser degree “Romance), and the actor S has a high score compared with other actors.

Online Recommendation Phase

In contrast to transaction-based usage profiles, semantic usage profiles capture the underlying common properties and relations among those objects. This fine-grained domain knowledge, captured in aggregate form enables more powerful approaches to personalization. As before, we consider the browsing history of the current user, i.e., active session, to be a weighted set of Web pages that the user has visited. The same transformation described in the last subsection can be used to create a semantic representation of the user’s active session. We call this representation the current user profile.

Figure 7 presents the basic procedure for generating recommendations based on semantic profiles. The recommendation engine matches the current user profile against the discovered domain-level aggregate profiles. The usage profiles with matching score greater than some pre-specified threshold are considered to represent this user’s potential interests. A successful match implies that the current user shares common interests with the group of users represented by the profile. The matching process results in an extended user profile which is obtained by applying the aggregation process described above to the domain-level profiles and the original user profile.

The recommendation engine then instantiates the user’s extended profile to real Web objects and will recommend them to the user. We can also exploit structural relationships among classes during the recommendation process. For example, if a concept hierarchy exists among objects, and the recommendation engine can not find a good match for a user profile at a certain concept level, then it can generalize to a more abstract level (e.g., from “romantic comedy” to “romance”).

This approach has several advantages over traditional usage-based personalization. First, it retains the user-to-user relationships that can be captured by the discovered usage profiles. Secondly, in contrast to standard collaborative
filtering, it provides more flexibility in matching aggregate usage profiles with the current user’s activity because the matching process involves comparison of features and relationships, not exact item identities. Thirdly, the items do not have to appear in any usage profiles in order to be recommended, since fine-grained domain relationships are considered during the instantiation process. The previous example shows that this approach can also be used to solve the “new item” problem. Furthermore, it can alleviate the notorious “sparsity” problem in collaborative filtering systems by allowing for “fuzzy” comparisons between two user profiles (or ratings). The basis for matching profiles does not have to be similar ratings on the same items. The comparison can be based on showing interest in different objects with similar properties (for example, purchasing items that have the same brand). Therefore, even if the raw transaction or rating data is sparse, the semantic attributes of items or users can be used to indirectly infer potential interest in other items.

CONCLUSION

We have explored various approaches, requirements, and issues for integrating semantic knowledge into the personalization process based on Web usage mining. We have considered approaches based on the extraction of semantic features from the textual content contained in a site and their integration with Web usage mining tasks and personalization both in the pre-mining and the post-mining phases of the process. We have also presented a framework for Web personalization based on full integration of domain ontologies and usage patterns. The examples provided throughout this chapter reveal how such a framework can provide insightful patterns and smarter personalization services.

We leave some interesting research problems for open discussion and future work. Most important among these are techniques for computing similarity between domain objects and aggregate domain-level patterns, as well as learning techniques to automatically determine appropriate combination functions used in the aggregation process.

More generally, the challenges lie in the successful integration of ontological knowledge at every stage of the knowledge discovery process. In the preprocessing phase, the challenges are in automatic methods for the extraction and learning of the ontologies and in the mapping of users’ activities at the clickstream level to more abstract concepts and classes. For the data mining phase, the primary goal is to develop new approaches that take into account complex semantic relationships such as those present in relational databases with multiple relations. Indeed, in recent years, there has been more focus on techniques such as those based relational data mining. Finally in the personalization stage, the challenge is in developing techniques that can successfully and efficiently measure semantic similarities among complex objects (possibly from different ontologies).
In this chapter we have only provided an overview of the relevant issues and suggested a road map for further research and development in this area. We believe that the successful integration of semantic knowledge with Web usage mining is likely to lead to the next generation of personalization tools which are more intelligent and more useful for Web users.

REFERENCES


Integrating Semantic Knowledge with Web Usage Mining for Personalization


This work was previously published in *Web Mining: Applications and Techniques*, edited by A. Scime, pp. 276-306, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 2.24
Modeling Variant User Interfaces for Web-Based Software Product Lines

Suet Chun Lee
BUSINEX, Inc., USA

ABSTRACT

Software product line (SPL) is a software engineering paradigm for software development. SPL is important in promoting software reuse, leading to higher productivity and quality. A software product within a product line often has specific functionalities that are not common to all other products within the product line. Those specific functionalities are termed “variant features” in a product line. SPL paradigm involves the modeling of variant features. However, little work in SPL investigates and addresses the modeling of variant features specific to UI. UML is the de facto modeling language for object-oriented software systems. It is known that UML needs better support in modeling UIs. Thus, much research developed UML extensions to improve UML support in modeling UIs. Yet little of this work is related to developing such extensions for modeling UIs for SPLs in which variant features specific to user interfaces (UI) modeling must be addressed. This research develops a UML extension, WUIML, to address these problems. WUIML defines elements for modeling variant features specific to UIs for Web-based SPLs. The model elements in WUIML extend from the metaclass and of the UML2.0 metamodel. WUIML integrates the modeling of variant features specific to UIs to UML. For example, in a Web-based patient registration SPL, member products targeting British users may use British date format in the user interface, while member products targeting United States users may use United States date format in the user interface. Thus, this is a variant feature for this product line. WUIML defines a model element, XOR, to represent such exclusive or conditions in a product line user interface model. WUIML would reduce SPL engineers’ efforts needed in UI development. To validate the WUIML research outcome, a case study was conducted. The results of this empirical study indicate that modeling UIs for Web-based SPLs using WUIML is more effective and efficient than using standard UML.
INTRODUCTION

Software product line (SPL) (Chastek, Donohoe, Kang, & Thiel, 2001; Clements & Northrop, 2002; SEI, 2005a) is a software engineering paradigm to develop software products. One important step in the SPL paradigm is the modeling of the functional features of software products across the product line. The features are called common core. An even more important step in the SPL paradigm is the modeling of the specific functional features within a particular member product in a product line. These specific functional features are called variant features because they are the features that differentiate member products in the product line. Then based on the model, a product is “assembled” by reusing the common core and selected variant features.

Unified Modeling Language (UML) (OMG, 2003b, 2004; Rumbaugh, Jacobson, & Booch, 2005) is a standard object-oriented modeling language. UML includes multiple views and diagram types to capture software functionalities from user perspective. However, UML seems to have not been developed for modeling user interface specific issues (Kovacevic, 1998; Silva & Paton, 2003). One of the usages of user interface models is that, in model-based user interface management systems (MB-UIMSs) (Griffiths et al., 2001; Szekey, Sukaviriya, Castells, Muthukumarasamy, & Salcher, 1996), user interface models can be used to generate user interface codes. There are extensions of UML (Blankenhorn & Jeckle, 2004; Nunes, 2003; Silva, 2002) to make UML better support user interface modeling. Yet, these extensions often assume the modeling of a single system instead of a SPL. On the other hand, although standard UML (OMG, 2003b, 2004) seems to have not been developed to support the modeling of SPLs, there are works (Gomaa, 2004; Gomaa & Gianturco, 2002; Ziadi, Héléouët, & Jézéquel, 2004) on extending UML to improve UML supports in modeling SPLs. Yet, these works do not focus on user interface modeling. Currently, many software products are Web-based. However, some (Silva, 2002) observe that there are specific modeling challenges for modeling user interfaces of Web-based software systems.

Thus, it is not clear how to model variant features for user interface specific issues in Web-based software product lines. This is an important barrier to overcome if software product line development of Web-based products is to take greater advantage of software reuse objectives: increased quality, decreased effort, or decreased time to market. Therefore, this paper is concerned with reporting research about developing a UML extension, Web User Interface Modeling Language (WUIML) that decreases effort by increasing effectiveness and efficiency needed in using UML to model user interfaces for Web-based software product lines. User interface development has been found (Myers, 1989) to account for a significant amount of overall software development work. WUIML would improve SPL software engineering paradigm in the user interface development perspective by reducing engineers’ efforts needed in user interface modeling.

BACKGROUND AND RELATED WORK

Unified Modeling Language

Unified Modeling Language (UML) (Booch, Rumbaugh, & Jacobson, 1999; OMG, 2003b, 2004; Scott, 2004) is a graphical language for specifying software systems. UML is a standard of the Object Management Group (OMG; see http://www.omg.org). The most current version for UML is UML 2.0 (OMG, 2003b, 2004). This research considers UML in UML 2.0 context.

UML is a standardized notation for object-oriented development. UML consists of views, diagrams, model elements, and general mechanisms. Views are used to present different aspects of complex systems from both the “system” in the
problem and the “system” in the solution. Tacitly, UML defaults to “system” as meaning the system in the solution. This is consistent with the classical use cases modeling and design process model. Therefore, whenever the term “system” is used in this paper, this usual UML use cases default as “system in the solution” is implied. Moreover, each view requires a number of diagrams, such as use case diagram, class diagram, and interaction diagrams. Each diagram captures a particular aspect of a system. Views that are often used with UML include: use case view, logical view, implementation view, process view, and deployment view. Use case views describe the functionality a system should deliver from external actors’ point of view. An actor is a human user or another system external to the system. A use case is often used to describe functions requested for a system. Logical views are used to describe a system’s functionality. Use cases are often the first set of models being created for a system. However, some found common pitfalls in use cases for one who is new to use case modeling (Lilly, 2000; Rosenberg & Scott, 2001).

Modeling concepts are captured in model elements. Example model elements are class, object, state, node, package, and component (Eriksson, Penker, Lyons, & Fado, 2004). Model elements are used in diagrams. A model element can have a visual representation. There are rules that limit types of model elements for different diagrams. Relationships between model elements are specified by relationships, such as association, generalization, dependency, and aggregation.

Additional information that cannot be represented using model elements is represented using UML general mechanisms, such as adornments, comments, tagged values, and properties (Eriksson et al., 2004). Adornments are used to add visual impact to the element. For example, “underline” an instance of a type is an adornment. Comments can be placed anywhere in a diagram. Comments often contain explanations or questions to resolve issues at a later time. Model elements have properties to store data values about an element. A property is defined with a name and a value called a tagged value. Properties are used to add additional specifications about element instances that are not normally shown in the diagram.

UML can be extended or adapted by its extension mechanisms: stereotype, tagged values, and constraints. A stereotype defines a new kind of model element based on existing model elements. A stereotype is described by placing its name as a string and within a pair of guillemots. For example, a class with the stereotype <<Radio>> is read as “a class of the Radio stereotype,” meaning that it is a radio type of class. The particular characteristics of a Radio class must have been defined when the stereotype is defined. A constraint is a restriction on an element that limits the usage of the elements. A constraint is enclosed in a curly bracket, for example, {student = “Dorothy”}. Alternatively, constraints can be written in Object Constraint Language (OCL) (OMG, 2003a). A UML profile (OMG, 2003b) is a stereotyped package that contains model elements that have been customized for a specific purpose using extension mechanisms, such as stereotypes, tagged values, and constraints. A profile can only extend an existing UML model element using stereotypes, tagged values, or constraints. A profile cannot create new model elements that were not previously defined in the UML 2.0 infrastructure specification (OMG, 2003b).

UML 2.0 is defined in Meta-Object-Facility (MOF) (OMG, 2002). MOF is an abstract language for defining modeling languages. Therefore, another way to extend UML is using MOF. MOF extends UML by defining new metamodel elements. This approach is often called metamodel approach. In metamodel approach, new model elements can be created. With metamodel approach, one can also create model elements that change UML symbols and semantics.
Software Product Line

Software product line (SPL) (SEI, 2005a) is a software engineering paradigm (SEI, 2005b). The paradigm mainly includes two software engineering processes: product line engineering and member product engineering. The product line engineering process involves a number of techniques in software engineering, such as domain analysis, requirements engineering, product architecture, and component development (Clements & Northrop, 2002). The product engineering process involves the configuration (“assemble”) of member products from the product line assets. The SPL paradigm depends on the notion of variability and variability mechanisms. In SPL, the term variability is used to indicate the extent of differences in user visible software properties among member products. A user visible software property is a feature. A feature of a member product that is different from other member products is a variant feature. Therefore, a variant feature is a feature only to one or more (but not all) products in a SPL. Variability mechanisms are a set of methods and steps that can be applied to develop variant features for SPLs.

The user interface of member products of a SPL may vary in one or more user interface specific features. Examples of user interface specific features are the layout of the user interface, the functionality of the user interface, the interaction of the user interface to and from the user, and so forth. The user interface specific variant features, like other non-user interface specific features, also need to be captured in a SPL model.

It seems that issues in user interface developments have been largely ignored by the SPL researches. It is not clear how to go about applying the software product line paradigm in user interface modeling. WUIML developed in this research is differentiated from other related works (Gomaa, 2004; Gomaa & Gianturco, 2002; Shin, 2002; Webber, 2001; Ziadi et al., 2004) in SPL researches in that WUIML is addressing the modeling of user interfaces in a SPL using UML. It is known (Silva & Paton, 2003) that UML needs better support in modeling user interfaces. The related works extend UML to improve support on modeling SPL, yet those works have not addressed the modeling issues in user interfaces for SPL. User interface development has been found (Myers, 1989) to account for a significant amount of overall software development work. WUIML would improve SPL software engineering paradigm in the user interface development perspective by reducing engineers’ efforts needed in user interface modeling.

Web User Interface Modeling

User interface is a computer-based media to facilitate communication between users and a software system (Marcus, 2002). Early user interfaces in software systems are text or form-based. Modern software systems however use graphical user interfaces (GUI) implemented according to programming packages, such as Java Swings classes (Eckstein, Loy, & Wood, 1998). More recently, HTML is used to create Web-based user interfaces that can be rendered by Internet browsers (Sommerville, 2001).

The user interfaces of a Web-based software product are called Web user interface (WUI) in this research. The modeling of a WUI is the development of an abstract understanding of the requirements of the WUI and represents it in an abstract notation. The modeling of WUI using WUIML does not provide a visual look of a WUI, thus, WUI modeling is not the design of a WUI. Currently, a major use of WUI modeling is to provide inputs for WUI code generation. In particular, WUI models resulting from WUI modeling are fed into a model-based tool. The tool either auto-generates (or semi-auto-generates) WUI implementations for the WUI models (Behrens, Ritter, & Six, 2003; Gómez & Cachero, 2003).
UML-Based User Interface Modeling

One approach to user interface modeling and specification methods is to extend UML notations (Behrens et al., 2003; Blankenhorn & Jeckle, 2004; Hennicker & Koch, 2001; Jacobson, Christerson, Jonsson, & Overgaard, 1992; Nunes, 2003; Phillips, Kemp, & Kek, 2001; Scogings & Phillips, 2001; Silva & Paton, 2003).

A Web UI development method is proposed in Behrens et al. (2003). The method consists of two tasks: UI requirements modeling and mapping UI requirements model to UI architecture. The modeling of UI requirements includes the specification of use cases (called “work units” in the paper), and then the work units are modeled using user-defined UML stereotypes representing user interface elements. A UI element is modeled as a \texttt{<<scene>>}. A \texttt{<<scene>>} may consist of zero or more \texttt{<<class view>>}. A user interface is then the composition of different \texttt{<<scene>>} elements. The navigations between user interfaces are based on the parent-child relationship between the user interfaces, domain requirements, and usability aspects. In Hennicker and Koch (2001), a UML profile for modeling user interfaces for Web applications is proposed. The UML profile includes UML stereotypes to model the presentation (“look and feel”) and navigation aspects of Web UIs. The Object-Oriented Hypermedia (OO-H) Method (Gómez & Cachero, 2003) extends UML with a set of new views for modeling Web user interface models. In Silva and Paton (2003), a UML extension, UMLi, was developed to support user interface modeling for interactive systems. However, UMLi has yet to address Web UI as indicated in Silva (2002) that there are additional user interface modeling properties for Web applications.

The major difference between this work and the related works (Behrens et al., 2003; Gómez & Cachero, 2003; Hennicker & Koch, 2001; Jacobson et al., 1992; Silva, 2002; Silva & Paton, 2003) in user interface modeling is that WUIML addresses the modeling of user interfaces for SPLs and the related works focus on the modeling of user interfaces for a single system. In addition, the modeling elements in WUIML capture the Web user interface properties specific to Web software products according to important Web standards (Axelsson, Epperson, Ishikawa, McCarron, Navarro, & Pemberton, 2003; Bos, Celik, Hickson, & Lie, 2004; Dubinko, Leigh, Klotz, Merrick, & Raman, 2002; Lie & Bos, 1999; Meyer, 2000; W3C). WUIML improves SPL software engineering paradigm by reducing efforts needed by SPL engineers in user interface modeling for SPLs.

WEB USER INTERFACE MODELING LANGUAGE (WUIML)

WUIML improves UML support on user interface modeling for Web-based software product lines. User interface modeling falls into the scope of user interface development. The user interface models specified in WUIML form one generic user interface software product line asset for the entire main software product line. The idea is to have user interface models as core assets. Then to develop user interfaces for member products, one would reuse all common user interface components and reuse selectively the variant user interface components.

In this research, the requirements analysis process for user interfaces is as follows. Given software requirements for a software product line, user interface related use cases from product line requirements are derived. A use case that is user interface related can be developed into many user interface scenarios. Each scenario is due to one or more variant user interface aspects. A user interface scenario involves one or more user interfaces. Any interaction between a human user and a user interface or between a user interface and a backend application is also captured in the use case scenario. Once the user interfaces are identified, one can decide on the user interface components. After eliciting and describing the scenarios, the
next step is to identify the relationships between user interfaces. The interacting relationships between user interfaces are captured in extended activity diagrams where nodes in the activity diagrams can represent variant user interfaces. The extended activity diagram shows the variant interactions between user interfaces.

In order to specify common and variant features for user interfaces, WUIML defines stereotyped relationships between modeling elements. These new relationships are selected syntheses and extensions from Jacobson, Griss, and Jonsson (1997) and Anastasopoulos and Gacek’s (2001) variability mechanisms. In addition, the variant notations and rules of the FODA (Cohen, Stanley, Peterson, & Krut, Jr., 1992) technique have been adopted and incorporated into WUIML. In addition, this research developed a new variability mechanism called WUIML. WUIDefinition models a WUI that is an aggregation of different parts of other WUIs. For example, a portal WUI is a WUIDefinition of a number of other WUIs. The difference between WUIDefinition and the UML aggregation is that WUIDefinition limits its aggregated components to be instances of WUIComposite or specific WUIElements while the standard UML aggregation does not have this restriction. This restriction helps modelers on deciding what components are appropriate for the aggregation. The variability mechanisms are used to generate variants on user interface aspects. User interface aspects include (but not limited to) user functional requirements, style, presentation, layout, events, data model, and constraints.

The structural contents of WUIML are built from various World Wide Web Consortium (W3C; see http://www.w3c.org) specifications: XHTML (Altheim & McCarron, 2001; Altheim, Boumphrey, McCarron, Schnitzendomer, & Wugofski, 2001; Altheim & McCarron, 2001; Axelsson et al., 2003; Powell, 2003; Raggett, Hors, & Jacobs, 1999; Sauers & Wyke, 2001; W3C, 2002), XForms (Dubinko, 2003; Dubinko et al., 2002; Khare, 2000), and Cascading Style Sheets (CSS) (Bos et al., 2004; Lie & Bos, 1999; Meyer, 2000, 2003; Schengili-Roberts, 2004; Schmitt, 2003).

WUIML is also built on the UML 2.0 specifications (OMG, 2003b, 2004) and is an extension to the metamodeling approach of the UML 2.0. In particular, WUIML extends metaclass and BasicActivity of the UML 2.0 metamodel. The metaclass extension is achieved via stereotype extension mechanism. The extension to the BasicActivity is achieved via MOF.

A WUI structurally consists of multiple user interface elements, such as user interface visual elements (such as a button, a checkbox, etc.), hypertext links, images, text, and so forth. These elements are “mapped” to stereotyped UML classes that extend the UML2.0 metaclass.

Events (sometimes termed “triggers”) occur when user interface elements interact with user actions. Actions are mapped to UML interfaces. An interface defines a set of operations triggered by events. Events and actions characterize a user interface interaction and are modeled as operations of metamodel elements. Interactions between user interfaces are modeled using UML activity diagrams. In addition, UML activity diagram is extended to model variant interactions in WUIML.

Basic WUIML Elements

Figure 1 defines the Basic WUIML Elements in UML 2.0 notations. These model elements extend from the metaclass of the UML2.0 metamodel. These model elements have attributes and operations. The attributes are used to model the static features of user interfaces; for example, an attribute can be used to describe the background color of a user interface element. The operations are used to model the dynamic aspects of user interfaces, for example, events (e.g., a user clicks a button can be tracked by an event listener) and actions (e.g., an error message box popup can be implemented by a method) can be modeled.
Modeling Variant User Interfaces for Web-Based Software Product Lines

Figure 1. Basic WUIML elements

as operations. Figure 2 shows the class diagram of the WUIElement and the Command model elements.

A WUI is modeled as a WUITemplate. A WUITemplate may be composed of one or more of WUIComposite. A WUITemplate may also be composed of one or more WUIElement and one or more WUIComposite. A WUIComposite has one or more specific WUI modeling elements (such as Submit, Input, etc.) that extend from the WUIElement. The style, layout, and position within a WUI of a WUIElement can be described using the StyleDecorator model element. A StyleDecorator composes of a number of Dimension element, Position element, and the Style element. The Dimension (how large or small the WUI element is), Position (where the WUI element is positioned within a WUI), and Style (such as what is the foreground color of the WUI element) elements together model abstractly the style of a WUI element. A set of StyleDecorator elements forms a StyleStrategy. A StyleStrategy element describes the style of a WUIComposite element. A set of StyleStrategy elements models the style of a WUITemplate.

These WUIML elements integrate Web properties according to important Web standards (Axelsson et al., 2003; Bos et al., 2004; Dubinko et al., 2002; Lie & Bos, 1999; Meyer, 2000; W3C) to the modeling of Web-based user interfaces for
SPLs. With regard to the modeling of the static aspects of user interfaces, what WUIML adds to the standard UML is like what application programming interfaces (APIs) add to a programming language. For example, the Swing API (Eckstein et al., 1998) provides the building blocks for programmers to code user interfaces. Similarly, WUIML provides the building blocks for modelers to model user interfaces. When using Swing, a programmer does not have to construct fundamental user interface elements, such as textboxes, radio buttons, and so forth. The programmer would just use them. Similarly, when using WUIML, a modeler does not have to model the fundamental user interface elements, such as textboxes, radio buttons, and so forth. When using standard UML, those fundamental user interface elements must be modeled first. In addition, WUIML provides a way to conveniently model the layout, size, and position (where it is within a user interface) for a particular user interface element. Currently, there is no standard way in using UML to model the layout and style of a Web-based user interface. WUIML defines a way for modeling the layout and style of a Web-based user interface.
Example I: Web User Interface Modeling Using Basic WUIML Elements

In a SPL, a user interface feature may vary among the member products. For example, consider a SPL of online patient registration systems. A member product to be deployed for U.S. customers would have a user interface customized for the U.S. address format. On the other hand, member products to be deployed for Hong Kong customers would have a user interface customized for the Hong Kong address format.

Figure 3 shows a product line Web user interface model that consists of instances of <<WUIComposite>> elements. An instance of <<WUIComposite>> may compose of a number of fundamental user interface elements (such as a textbox, an input box, a listbox, etc.) but they are not shown in Figure 3. A <<WUIComposite>> can be considered as a portion of a user interface called user interface fragment in this research. A user interface is composed of a number of user interface fragments.

Modeling Variant Interactions

WUIML defines elements to model the different interactions between a user interface and its back-end process. If an MVC architecture (Gamma, Helm, Johnson, & Vlissides, 1995) is the example, then the back-end application is the controller and the model. For the same functional task, such as a WUI for remittance of an online banking service SPL, each member product may implement it in a slightly different way. The difference would lead to differences in the WUI related to the task among member products. The differences in WUI would in turn lead to different patterns of user interactions to the WUI as well as different implementation of back-end processes. For example, to send money to a foreign country, the WUI must collect the destination country (as country code) and currency (currency code) information from the user. On the other hand, to send money...
domestically within the U.S., the WUI does not need to collect country code or currency information. The different user interactions to a user interface in a SPL are called variant interactions. User actions, such as clicking a submit button in a WUI, may trigger transitions between WUls. In this research, UML activity diagram is used to model the interactions between WUls.

WUIML defines two new metamodel elements using MOF to model variant interactions. The two new elements are VariantAction and SPLAction. In UML, a node in an activity diagram is represented by the ActivityNode element. The ActivityNode element has three children: ExecutableNode, ObjectNode, and ControlNode. Object Nodes represent the objects involved in the activity diagram. The ControlNode signifies the control flow on the activity diagram; for example, a fork bar is a ControlNode. The ExecutableNode has a child, the Action element. The Action element in UML is the representation of a node that characterized by one or more actions. Since interactions are characterized by actions, WUIML extends the UML Action element to model variant interactions.

Figure 4 shows the new metamodel elements, VariantAction and SPLAction, and how they relate to other elements in the activity class diagram. The VariantAction element is a specialization of the UML Action element. The SPLAction is a specialization of the Element metamodel element from the UML2.0 kernel package. A SPLAction element contains one or many Variant Action elements. A UML Activity is now also consisting of one or more of the SPLAction and VariantAction elements.

Suppose we have an activity called “submit order.” The “submit order” activity is very common in Web-based shopping applications because a user must submit an order so that the order can be processed. But there are variants in the “submit order” activity. For example, most online shopping sites accept only credit card payments, thus, those applications may want the “submit credit
Modeling Variant User Interfaces for Web-Based Software Product Lines

In another case, corporate purchasing often submits purchase orders, and then pays the bill after receiving the invoice. In this case, “submit purchase order” may be desired. Yet another variant is in the case where a customer wants to receive the purchase and check it first before making a payment, these customers may want the “submit Cash-On-Delivery (COD) order.” Yet some merchants may offer a combination of these payment methods. There are a number of variants in the “submit order” activity, and we need to be able to model the variants. This paper extends UML activity diagram to model the requirements as follows.

Figure 5 shows a graphical symbol for the SPLAction. The symbol exposes the base action and its four variants actions. The shape of the base “submit order” action is filled, while the variants are not. The base Submit Order action defines the common logics that are applicable across all of the variants. Each variant differs from other variants by some features. Our concern is on user interface modeling; it is clear that the user interface to collect information on a credit card order is different from the user interface to collect a COD order in some way.

Sometimes, there are variant logics and behaviors within a variant member product that capture different execution scenarios. Those logics and behaviors can be modeled using standard UML activity diagrams. For example, suppose there is a WUI that provides user interface elements to accept either credit card based order option or purchase order option. If a customer chooses credit card order, the next WUI displayed is to collect the customer’s credit card information. If a customer chooses purchase order based order option, then the next WUI displayed is to collect the customer’s purchase order related information. Thus, the software’s behavior varies due to different execution scenarios. This variant is within a product.

These variant logics and behaviors, due to different execution scenarios, often are easy for one to confuse them with the product line variant logics and behaviors. To identify the product line variants, one must consider the same particular variant logics and behaviors across multiple member products at the same time. If the variant logics and behaviors work the same across the member products, then the variant logics and behaviors are not a product line variant.

Example II: Modeling Variant Interactions

Figure 6a shows an activity diagram for a SPLAction for an online e-retailer SPL. The SPLAction is about “Submit Order.” The SPLAction consists of four variant actions: “Submit Credit Card Order,” “Submit COD Order,” “Submit Check Order,” and “Submit Purchase Order.” A member product of this product line would have the same activity but with only one variant of the submit order action. Figure 6b shows the activity diagram for a member product that allows users to submit Credit Card Orders.

WUIML Elements for Variability Mechanisms

Figure 7 shows the WUIML elements for variability mechanisms that are based on Jacobson, Griss, and Jonsson (1997) and Anastasopoulos and Galek’s (2001) variability mechanisms and the variant notations and rules of the FODA (Cohen
Modeling Variant User Interfaces for Web-Based Software Product Lines

Figure 6a. Exemple SPL variant interactions

Figure 6b. Example of a member product variant interaction

Figure 7. Model elements for variability mechanisms
et al., 1992) technique. Delegation, WUIDerive, WUIExtend, Use, TemplateInstantiation, and RequirePresenceOf are stereotypes extended from Dependency metaclass. WUIAggregation, Configuration, and Parameterization are defined as stereotypes of Class stereotype. WUIGeneralization is a stereotype of the Generalization metaclass. OptionalElement is a constraint to express that an element is optional in the model. XOR and OR are Constraints. A Parameterization element is composed of one or more Property. A Configuration element is composed of one or more Property and zero or more Constraint.

The Use model element is used to specify that the source user interface element (e.g., a user interface fragment to capture the travel date) depends on the target user interface element (e.g., a user interface of a calendar for setting a travel date). The Configuration model element models a set of parameters and rules. These parameters and rules manipulate the composition of the software component of a software system or the setting of software components or software systems, so that variants can be created. The Parameterization model element models a set of parameters that vary the features of a software product. The TemplateInstantiation model element is a special case of the Parameterization model element. In TemplateInstantiation, the parameters are templates. WUIExtend allows small extensions in functionality or “look” to a user interface due to new requirements. WUIDerive extends UML derive stereotype to apply on all basic WUIML elements. Delegation is used to model the situation in which a Web user interface is acting as a representative of another Web user interface in performing certain operations. WUIAggregation allows one to model the situation in which various user interface fragments from various Web user interfaces are used to compose another Web user interface based on a business goal. WUIAggregation differs from UML aggregation in that the aggregated components in WUIAggregation must be an instance of WUIComposite or specific WUI elements. WUIGeneralization models the “is-a-kind-of” relationship between two user interfaces. WUIGeneralization differs from UML generalization in that the generalization in WUIGeneralization applies to WUI elements only. OptionalElement is used to specify that a user interface element is optional in the model. OR models the inclusive or constraint. XOR models the exclusive or constraint. RequirePresenceOf models the situation in which a user interface feature depends on the presence of another user interface feature. For example, to model the style of a column within a table, it requires the presence of a table element.

**Example III: Modeling Variant Web User Interface Features**

To illustrate WUIML for modeling variant WUI features, let us continue the example shown in Figure 3. But this time, we apply the variability mechanisms using the corresponding variability mechanism model elements defined in WUIML. Figure 8 shows a more detailed WUIML model for the online patient registration systems SPL. Figure 3 shows the common user interface features across the product line. Figure 8 shows the variant user interface features among member products within the product line in addition to the common features across the product line.

Applying variability mechanisms and the corresponding WUIML notations, variant WUI features are specified in Figure 8. The variant notations in WUIML are only used in product line WUIML models to show the variant features among member products. Thus, the variant notations in WUIML are not to appear in the WUIML model of a member product. For example, notice that the USAddress and HKAddress WUIComposite elements appear in Figure 3 are now specified with exclusive or constraint. This indicates that if a member product’s WUI uses USAddress format, then it will not use the HKAddress format. TemplateInstantiation vari-
ability mechanism is used to create two variant WUIComposites of PatientInfoReporter. The patient information report can be in graphic format or summary format.

**Modeling Dynamic Aspects of Web User Interfaces**

In addition to interactions as described in the section *Modeling Variant Interactions*, the dynamic aspects of WUIs also include events and actions. In this paper, event literally means “something that takes place; an occurrence” (Pickett et al., 2000). In this paper, an event means an occurrence that takes place in the user interface. Example occurrence may be a mouse click on a button, a selection on a selection box, and so forth. The term action literally means “the state or process of acting or doing” (Pickett et al., 2000). In this paper, an action means doing some user interface related tasks in response to user interface events. An example action may be popping up a message box on the user interface, setting a value for a text input box, and so forth. Moreover, an event is observed by a user interface control, such as a button, and then the user interface control triggers an action. For example, when a button is clicked, a message box is popped up to display a message. In this scenario, the button is the user interface control that is observing the event: clicking the button. The popping up of a message box is the action being triggered by the button click event. The concept of events and actions described in this section are built on the XForms specification.

XForms specification includes a set of events and actions. Many events and actions have to do with XForms processing and are not suitable for use at modeling level. For this reason, only a selected set of events and actions are taken as the bases for building WUIML elements for modeling the dynamic nature of a user interface. In WUIML, events are modeled as operations of a basic WUIML metamodel element. Each basic WUIML metamodel element has a distinct set
of events. Extra events can be added as needed by the modeler at design time as extensions. An action models the resulting behavior of a basic WUIML element in response to events. Actions are represented as operations of a basic WUIML element.

**Example IV: Modeling Product Web User Interface Using WUIML**

Figure 9 shows an even more detailed WUIML model for the online patient registration systems SPL than Figure 8. In Figure 9, more variability mechanisms are applied to create variants. In addition to the variability mechanisms shown in Figure 8, the variability mechanisms applied in Figure 9 include WUIExtend, Use, Configuration, WUIAggregation, Delegation, and WUIDerive. Once we have a SPL WUIML model, we can derive product WUIML model from the SPL WUIML model. A product WUIML model is a WUIML model that reuses all the common features of a SPL WUIML model but reuses only selected variant features from a SPL WUIML model. Figure 10 and Figure 11 show two product WUIML models derived from Figure 9.

The variant patient registration WUIML model shown in Figure 10 is created by taking all the common model elements (i.e., Email, LoginInfo, Address, PhoneNum, NameInfo, OtherContent, PatientInfoReporter, and SubmitRequest) and selected variant user interface features (i.e.,...
USAddress, USPhoneNum, and NameInfo_US-Style) from Figure 9.

Figure 11 shows another variant product WUIML model derived from Figure 9. Compare Figure 11 with Figure 10; Figure 11 shows more variant user interface features. Figure 11 includes the BritishDate WUIComposite that is extended from DateRegistered WUIComposite.

VALIDATION

The goal of this research is to improve UML support in modeling user interfaces for Web-based SPLs. The improvement goal is to have the new WUIML method exhibit decreased effort needed by increasing effectiveness and efficiency. The improvement will reduce the SPL engineers’ efforts needed for developing user interfaces. To exhibit that this improvement goal is met, a case study research validation method is applied to supply supporting evidence for the thesis (research hypothesis).

Case Study Research Method

Using case study (Lynch, 1999; Tellis, 1997a, 1997b; Yin, 2003) as an empirical research validation method has increasingly been adopted in software engineering research (Lee, 2004). This research uses a case study research method to validate the research hypothesis.

The rationale of a case study research method is the notion of “analytical generalization.” Analytical generalization depends on whether the findings of a case study can be replicated. If the findings of a case can be replicated, an analytical generalization may be drawn.

In analytical generalization, the results of a case study should be compared with a proposition. For example, in this case study, the results of a multiple-case study are to compare with proposition 1 and proposition 2 (proposition 1 and proposition 2 will be discussed later). Multiple cases are needed to test a proposition through replications of findings in order to lead to an analytical generalization (Yin, 2003). Many research methods, such as a survey, depend on statistical generalization (Yin, 2003). In statis-
cal generalization, an inference is made about a population (or universe) base on the empirical data collected about a sample (i.e., surveys) (Yin, 2003). Since a case study is not based on statistical generalization, a sampling logic should not be used in a case study research method. As a result, the number of cases in a multiple-case design and the typical criteria regarding sample size are irrelevant (Yin, 2003).

The evidence from multiple-case studies makes the overall study more robust (Yin, 2003). In this empirical study, multiple case studies were conducted. In a multiple-case study, each case must be selected to predict similar results (a literal replication) or to produce contrasting results but for predictable reasons (a theoretical replication) (Yin, 2003). A literal replication explains the conditions under which a particular phenomenon is likely to be found. A theoretical replication explains the conditions when it is not likely to be found. This empirical study relies on literal replications. The two case studies conducted are expected to predict similar results. The study is trying to show that under the condition (i.e., using WUIML in modeling WUIs for a SPL), a particular phenomenon (i.e., increased modelers’ effectiveness and efficiency in modeling) is likely to be found. For each case in the multiple-case study, the study indicates the way a particular proposition was demonstrated (or not demonstrated). Across cases in the multiple-case study, the study indicates the extent of the replication logic and the reasons on the prediction of supporting or contrasting results (Yin, 2003).

**Case Study Design**

The goal of this empirical study is to investigate the effectiveness and efficiency of WUIML. This empirical study uses a case study research method (Lee, 2004; Tellis, 1997a, 1997b; Yin, 2003) as the validation method. The design of a case study is characterized by five important case study components (Yin, 2003): a study’s research
questions, study propositions, units of analysis, the logic linking data (results) to the propositions, and the criteria for interpreting the findings.

Study’s Research Questions

The study’s research questions define the validation goal of a case study. The study’s research questions should be clarified precisely (Yin, 2003). For example, in this empirical study, the study’s research question is “Does Web user interface modeling for a Web-based medical SPL using WUIML increase SMEs’ modeling efficiency and effectiveness (thus decreasing work)?” This study’s research question needs to be clarified further because the notion of efficiency and effectiveness need further clarification. This study’s research question can be decomposed into a set of propositions.

Study Propositions

Study propositions are derived from the study’s research questions (Yin, 2003) but are more specific than the study’s research questions. Study propositions quantify the quality variables (indirect metrics) in a study’s research question into directly measurable quantitative metrics (direct metrics or indicators). For example, in this multiple-case study, the study’s research question is decomposed into two propositions:

1. SMEs are able to correctly model a larger number of required modeling items using WUIML than standard UML in modeling WUIs for a Web-based medical SPL in approximately the same amount of person-hours.

2. SMEs are able to correctly model larger numbers of required modeling items via reuse using WUIML than standard UML in modeling WUIs for a Web-based medical SPL in approximately the same amount of person-hours.

In this multiple-case study, for each WUI, the investigator has identified from the requirements a set of modeling items, called required modeling items that must be modeled by the SMEs. The resulted models produce by SMEs are inspected by the investigator. The investigator first checks the models for required modeling items. Then, based on his modeling experience, the investigator decides the correctness of the required modeling items found in the resulted models.

Units of Analysis

Units of analysis are materials such as documents or other resources that the subject matter experts (SMEs) use as inputs or materials to apply the method or tools being validated. In this study, the method under investigation is WUIML. The units of analysis are the Web user interface requirements for a medical SPL.

In the first case study, the requirements for the Pediatric Medical Profile Login WUI, the Adolescent Medical Profile Login WUI, and the Adult Medical Profile Login WUI are provided. These WUIs are each from a different member product (Pediatric Medical Management System, Adolescent Medical Management System, and Adult Medical Management System) of a medical product line (Medical Management Systems). The WUIs to model are extracted from three Web-based medical products under development in BUSINEX Inc.

The WUIs for the medical Web software products are based on actual medical forms from health-care providers in the United States. For the first case study, three member products of the product line are considered for WUI modeling. In particular, this case study requires the SMEs to model the Medical Profile Login WUI and a related activity across three member products of the product line. This WUI is chosen because it allows one to exercise the modeling of commonality and variability found in product lines.
In the second case study, the requirements for the Pediatric Exercise Record WUI, the Adolescent Exercise Record WUI, and the Adult Exercise Record WUI are provided. These WUIs are each from the same member products in the same product line as in the first case study. These requirements are the units of analysis for the second case study.

Note that within each case study, there are two SMEs. One SME applies WUIML to the requirements to generate results while the other SME applies standard UML to the requirements to generate results. The results generated by applying standard UML are used as the baseline for analyzing the results generated by applying WUIML.

In this research, the SMEs are well trained in software engineering; they are professional software engineers with experience ranging from 7 to 15 years. Their specialties are focused on user interface development. They are representative users for the new WUIML. The investigator provides training on WUIML to the SMEs before they begin conducting the case studies. In each case, all SMEs participate in the case study only once. Not having the same SME to participate in

<table>
<thead>
<tr>
<th>Required modeling items for Pediatric Medical Profile Login WUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page title: “Pediatric Medical Profile Login”</td>
</tr>
<tr>
<td>Label 1: “Profile ID:”</td>
</tr>
<tr>
<td>A textbox for label 1.</td>
</tr>
<tr>
<td>Label 2: “Password:”</td>
</tr>
<tr>
<td>A textbox for label 2.</td>
</tr>
<tr>
<td>Label 3: “Role:”</td>
</tr>
<tr>
<td>A radio button.</td>
</tr>
<tr>
<td>The radio button must default to be checked.</td>
</tr>
<tr>
<td>A label for the radio button: “Parent/Guardian”. (Only parent or legal guardian who are previously registered with the health provider can login for the child.).</td>
</tr>
<tr>
<td>A submit button with name “Login”.</td>
</tr>
<tr>
<td>The Page title must placed on the top of the page.</td>
</tr>
<tr>
<td>The Profile ID and its related textbox must be immediately next to each other.</td>
</tr>
<tr>
<td>The Password and its related textbox must be immediately next to each other.</td>
</tr>
<tr>
<td>The Role and its related radioButton must not be immediately next to each other.</td>
</tr>
<tr>
<td>Profile ID must be placed on top of the Password and its related textbox.</td>
</tr>
<tr>
<td>The Role and the radio button must be on a line that is on top of the Login button.</td>
</tr>
<tr>
<td>The Login button must be placed at the lower left hand side of the page.</td>
</tr>
<tr>
<td>The activity diagram should include an action node: Pediatric Medical Profile Login.</td>
</tr>
<tr>
<td>The activity diagram should include an action node: Login Error.</td>
</tr>
<tr>
<td>The activity diagram should include an action node: Parent Welcome.</td>
</tr>
<tr>
<td>The activity diagram should include an action node: Customer Service.</td>
</tr>
<tr>
<td>The activity diagram should include a start node.</td>
</tr>
<tr>
<td>The activity diagram should include an end node.</td>
</tr>
<tr>
<td>The activity diagram should include a decision node: whether validation is successful.</td>
</tr>
<tr>
<td>The activity diagram should indicate the condition that “Profile ID, Password, and Role” values are available.</td>
</tr>
</tbody>
</table>
more than one case is to prevent the introduction of bias due to the familiarity of WUIML by SMEs participating in multiple case studies. The SMEs are to carry out the case study without help from the investigator. The SMEs are also not given information on the expected results of the case study. This is to prevent bias by the SMEs in performing the case studies.

Table 1 shows the required modeling items for the Pediatric Medical Profile Login WUIs. The result of whether these items are correctly modeled by SMEs, the number of correctly modeled items over total number of required modeling items by SMEs with or without product line reuse, and the approximate person-hours spent will be used to support/reject the study propositions 1 and 2. There are specific required modeling items for Adult Medical Profile Login WUI, Pediatric Medical Profile Login WUI, Adult Exercise Record WUI, Adolescent Exercise Record WUI, and Pediatric Exercise Record WUI; for brevity, they are not shown here.

**Linking Data to Propositions**

Data analysis is done in the step of linking data (results) to propositions. In a case study, data are collected using formal case study worksheets. A case study worksheet consists of specific concrete questions associate with each one of the study propositions. The answers to those questions are measured to concrete criteria metrics that can

---

### Figure 12. Case study worksheet for modeling Pediatric Medical Profile Login WUI using WUIML

#### Case Study Worksheet

Using WUIML to model the Pediatric Medical Profile Login WUI.

**SME ID:**

Please answer the following questions:

1. **What tool** did you use to perform the modeling?

2. **What is the filename of the resulted model(s)?**

3. **Please list below the model elements** (e.g. classes and activity nodes, etc.) in your resulted models. For each element, first provide its name and then provide a brief description on its purpose. List as many as you wish in additional pages (if needed).

4. **Among the items listed in 3, which of the model elements are developed via reuse?**

5. **How many person-hours** did you spend in complete the modeling of this WUI?
be analyzed to support (or reject) the propositions. Thus, the most concrete criteria metrics are measured terms found within the questions on the case study worksheets. Figure 12 shows a case study worksheet used in the first case study for modeling the Pediatric Medical Profile Login WUI. In Figure 12, each concrete criteria metrics in questions that link to propositions is identified by a name formed by three sections. For example, “Pediatric_WUIML_personHours” is the metric about person-hours spent in using WUIML to model the Pediatric Medical Profile Login WUI. The same format is used in naming other metrics found in the questions in other case study worksheets in the case studies.

Table 2 summarizes the evidence collected through the case study worksheets in the case study about Medical Profile Login WUIs.

In each case study, WUIML and standard UML are applied to the units of analysis (that is

<table>
<thead>
<tr>
<th>Questions (Units)</th>
<th>Evidence captured</th>
<th>Propositions to support/reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pediatric_WUIML_modelingTool</td>
<td>The modeling tool used to perform the modeling in WUIML for the Pediatric Medical Profile Login WUI</td>
<td>Proposition 1 and Proposition 2</td>
</tr>
<tr>
<td>Pediatric_WUIML_modelsFilename</td>
<td>The name of the software copy of the resulted models in WUIML for the Pediatric Medical Profile Login WUI</td>
<td>Proposition 1 and Proposition 2</td>
</tr>
<tr>
<td>Pediatric_WUIML_required_modeling_item</td>
<td>The correctly modeled required modeling items in WUIML for the Pediatric Medical Profile Login WUI</td>
<td>Proposition 1</td>
</tr>
<tr>
<td>Pediatric_WUIML_reuse</td>
<td>The correctly modeled required modeling items via reuse in WUIML for the Pediatric Medical Profile Login WUI</td>
<td>Proposition 2</td>
</tr>
<tr>
<td>Pediatric_WUIML_personHours</td>
<td>The person-hours spent to model the Pediatric Medical Profile Login WUI using WUIML</td>
<td>Proposition 1 and Proposition 2</td>
</tr>
<tr>
<td>Pediatric_UML_modelingTool</td>
<td>The modeling tool used to perform the modeling in UML for the Pediatric Medical Profile Login WUI</td>
<td>Proposition 1 and Proposition 2</td>
</tr>
<tr>
<td>Pediatric_UML_modelsFilename</td>
<td>The name of the software copy of the resulted models in UML for the Pediatric Medical Profile Login WUI</td>
<td>Proposition 1 and Proposition 2</td>
</tr>
<tr>
<td>Pediatric_UML_required_modeling_item</td>
<td>The correctly modeled required modeling items in UML for the Pediatric Medical Profile Login WUI</td>
<td>Proposition 1</td>
</tr>
<tr>
<td>Pediatric_UML_reuse</td>
<td>The correctly modeled required modeling items via reuse in UML for the Pediatric Medical Profile Login WUI</td>
<td>Proposition 2</td>
</tr>
<tr>
<td>Pediatric_UML_personHours</td>
<td>The person-hours spent to model the Pediatric Medical Profile Login WUI using UML</td>
<td>Proposition 1 and Proposition 2</td>
</tr>
<tr>
<td>Adolescent_WUIML_modelingTool</td>
<td>The modeling tool used to perform the modeling in WUIML for the Adolescent Medical Profile Login WUI</td>
<td>Proposition 1 and Proposition 2</td>
</tr>
<tr>
<td>Adolescent_WUIML_modelsFilename</td>
<td>The name of the software copy of the resulted models in WUIML for the Adolescent Medical Profile Login WUI</td>
<td>Proposition 1 and Proposition 2</td>
</tr>
</tbody>
</table>

continued on following page
**Modeling Variant User Interfaces for Web-Based Software Product Lines**

**Table 2. (continued)**

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Proposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adolescent_WUIML_required_modeling_items</td>
<td>The correctly modeled required modeling items in WUIML for the Adolescent Medical Profile Login WUI.</td>
<td>1</td>
</tr>
<tr>
<td>Adolescent_WUIML_reuse</td>
<td>The correctly modeled required modeling items via reuse in WUIML for the Adolescent Medical Profile Login WUI.</td>
<td>2</td>
</tr>
<tr>
<td>Adolescent_WUIML_personHours</td>
<td>The person-hours spent to model the Adolescent Metrical Profile Login WUI.</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Adolescent_UML_modelingTool</td>
<td>The modeling tool used to perform the modeling in UML for the Adolescent Medical Profile Login WUI.</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Adolescent_UML_modelsFilename</td>
<td>The name of the software copy of the resulted models in UML for the Adolescent Medical Profile Login WUI.</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Adolescent_WUIML_required_modeling_items</td>
<td>The correctly modeled required modeling items in WUIML for the Adolescent Medical Profile Login WUI.</td>
<td>1</td>
</tr>
<tr>
<td>Adolescent_UML_reuse</td>
<td>The correctly modeled required modeling items via reuse in UML for the Adolescent Medical Profile Login WUI.</td>
<td>2</td>
</tr>
<tr>
<td>Adolescent_UML_personHours</td>
<td>The person-hours spent to model the Adolescent Metrical Profile Login WUI.</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Adult_WUIML_modelingTool</td>
<td>The modeling tool used to perform the modeling in WUIML for the Adult Medical Profile Login WUI.</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Adult_WUIML_modelsFilename</td>
<td>The name of the software copy of the resulted models in WUIML for the Adult Medical Profile Login WUI.</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Adult_WUIML_required_modeling_items</td>
<td>The correctly modeled required modeling items in WUIML for the Adult Medical Profile Login WUI.</td>
<td>1</td>
</tr>
<tr>
<td>Adult_WUIML_reuse</td>
<td>The correctly modeled required modeling items via reuse in WUIML for the Adult Medical Profile Login WUI.</td>
<td>2</td>
</tr>
<tr>
<td>Adult_WUIML_personHours</td>
<td>The person-hours spent to model the Adult Metrical Profile Login WUI.</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Adult_UML_modelingTool</td>
<td>The modeling tool used to perform the modeling in UML for the Adult Medical Profile Login WUI.</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Adult_UML_modelsFilename</td>
<td>The name of the software copy of the resulted models in UML for the Adult Medical Profile Login WUI.</td>
<td>1 and 2</td>
</tr>
<tr>
<td>Adult_UML_required_modeling_items</td>
<td>The correctly modeled required modeling items in UML for the Adult Medical Profile Login WUI.</td>
<td>1</td>
</tr>
<tr>
<td>Adult_UML_reuse</td>
<td>The correctly modeled required modeling items via reuse in UML for the Adolescent Medical Profile Login WUI.</td>
<td>2</td>
</tr>
<tr>
<td>Adult_UML_personHours</td>
<td>The person-hours spent to model the Adult Metrical Profile Login WUI using UML.</td>
<td>1 and 2</td>
</tr>
</tbody>
</table>
Modeling Variant User Interfaces for Web-Based Software Product Lines

the Web user interface requirements of a Web-based medical SPL) respectively. The results, that is, the resulting models and the completed use case worksheets, are analyzed to find out the following:

D1. How many of the required modeling items are correctly modeled when the modeling was done in WUIML?

D2. How many of the required modeling items are correctly modeled when the modeling was done in standard UML?

D3. How many of the required modeling items are correctly modeled via reuse when the modeling was done in WUIML? To model via reuse is to create new models by re-using previously created models or model elements. For example, suppose one previously created a model that includes a class representing fruit. Now one can reuse the fruit class to create a class that represents a specific fruit, such as an apple, by extending the fruit class. Both standard UML and WUIML allow SMEs to model via reuse.

D4. How many of the required modeling items are correctly modeled via reuse when the modeling was done in standard UML?

D5. How many person-hours spent to generate the WUIML models?

D6. How many person-hours spent to generate the standard UML models?

D7. The total number of required modeling items. D1, D2, D5, D6, and D7 link to proposition 1. D3, D4, D5, D6, and D7 link to proposition 2.

Criteria for Interpreting a Case Study’s Findings

In this empirical study, the criteria for interpreting a case study’s findings correspond to the metric and measures used in evaluating the results of applying WUIML and standard UML respectively to the units of analysis. “A measure provides a quantitative indication of the extent, amount, dimension, capacity, or size of some attribute of a product or process” (Pressman, 2005). A metric is an important, directly-measurable attribute of a software product, a software service, a software process or a software resource. Direct metrics may be either used as indicators (predictors), or of other more valuable outcomes that are indirect metrics. For example, the indirect metrics in this case study are efficiency and effectiveness. The direct metrics are:

m1. Number of the required modeling items that are correctly modeled when the modeling was done in WUIML;

m2. Number of the required modeling items that are correctly modeled when the modeling was done in standard UML;

m3. Number of the required modeling items that are correctly modeled via reuse when the modeling was done in WUIML;

m4. Number of the required modeling items that are correctly modeled via reuse when the modeling was done in standard UML;

m5. Person-hours spent to generate the WUIML models;

m6. Person-hours spent to generate the standard UML models;

m7. The total number of required modeling items.

The Case Study Procedure for Both Case Studies

0. For those SMEs that need to perform the modeling in WUIML, teach them WUIML.

1. Identify the required modeling items from the software requirements.

2. Provide SMEs who are to conduct the modeling using WUIML with the software requirements.

3. Provide SMEs who are to conduct the modeling using standard UML with the software requirements.
4. Collect the results (WUIML models and the data from the completed case study worksheets) generated from step 2.
5. Collect the results (standard UML models and the data from the completed case study worksheets) generated from step 3.
6. Perform data analysis on the results and data collected from step 4 to find out the following: (a) number of the correctly modeled required modeling items when the modeling was done in WUIML (the measure of m1); (b) number of the correctly modeled required modeling items via reuse when the modeling was done in WUIML (the measure of m3); and (c) person-hours spent to generate the WUIML models (the measure of m5).
7. Perform data analysis on results and data collected from step 5 to find out the following: (a) number of the correctly modeled required modeling items when the modeling was done in standard UML (the measure of m2); (b) number of the correctly modeled required modeling items via reuse when the modeling was done in standard UML (the measure of m4); and (c) person-hours spent to generate the standard UML models (the measure of m6).
8. Evaluate the outcome from step 6 to determine whether or not the proposition 1 is supported or rejected.
9. Evaluate the outcome from step 7 to determine whether or not the proposition 2 is supported or rejected.

RESULTS AND ANALYSIS

WUIML is developed to improve SPL software engineering paradigm in modeling user interfaces for Web-based SPLs. The improvement provided by WUIML should decrease the work needed by the SPL engineers in the WUI modeling perspective.

In the first case study, each SME has to model the WUIs for three product lines: Pediatric Medical Profile Login, Adolescent Medical Profile Login, and Adult Medical Profile Login. One SME is asked to model each WUI using WUIML while the other SME is asked to model the WUIs using standard UML. Each WUI is from a different member product (Pediatric Medical Management System, Adolescent Medical Management System, and Adult Medical Management System) of a medical product line (Medical Management Systems).

Table 3 shows the results for modeling Pediatric Medical Profile Login WUI using standard UML and WUIML, respectively. The last row in Table 3 shows the ratio of the number of correctly modeled required modeling items to the total number of required modeling items in standard UML and WUIML, respectively.

Note that in each case study, there are two SMEs. One SME is asked to model using standard UML while the other SME is asked to model using WUIML. In the following tables (Tables 4, 5, 6, and 7), the numbers shown in the second column from the left were derived from the models created by the SME which was asked to model using standard UML. The numbers shown in the third column from the left were derived from the models created by the SME which was asked to model using WUIML. The sample size is irrelevant in case study research method because case study research method is based on analytical generalization instead of statistical generalization (Yin, 2003). However, the results shown in these tables must replicate (or be replicated by) the results from the second case study in order to lead to an analytical generalization.

Table 4 shows the ratio of the number of correctly modeled required modeling items to the total number of required modeling items in modeling the three WUIs and the person-hours spent on each approach. Notice that the person-hours spent for modeling the Pediatric Medical Profile Login WUI was more than the person-hours spent...
for Adolescent Medical Profile Login WUI and Adult Medical Profile Login WUI respectively in WUIML approach. To model the WUIs for a member product (such as the Pediatric Medical Profile Login WUI), one must first develop the WUI models for the SPL.

Table 5 shows the number of correctly modeled required modeling items per person-hours.
Modeling Variant User Interfaces for Web-Based Software Product Lines

Table 4. Ratio of correctly modeled required modeling items to the total number of required modeling items

<table>
<thead>
<tr>
<th>WUI</th>
<th>Standard UML (Person-hours)</th>
<th>WUIML (Person-hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pediatric</td>
<td>18/25 (8)</td>
<td>24/25 (7)</td>
</tr>
<tr>
<td>Adolescent</td>
<td>23/30 (10)</td>
<td>29/30 (4)</td>
</tr>
<tr>
<td>Adult</td>
<td>18/25 (7)</td>
<td>24/25 (2)</td>
</tr>
</tbody>
</table>

Table 5. Number of correctly modeled required modeling items

<table>
<thead>
<tr>
<th></th>
<th>Standard UML model</th>
<th>WUIML model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pediatric</td>
<td>2.25</td>
<td>3.4</td>
</tr>
<tr>
<td>Adolescent</td>
<td>2.3</td>
<td>7.25</td>
</tr>
<tr>
<td>Adult</td>
<td>2.57</td>
<td>12</td>
</tr>
<tr>
<td>Average</td>
<td>2.37</td>
<td>9.08</td>
</tr>
</tbody>
</table>

Table 6. Number of correctly modeled required modeling items via reuse

<table>
<thead>
<tr>
<th>WUI</th>
<th>Standard UML (Person-hours)</th>
<th>WUIML (Person-hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pediatric</td>
<td>0/10 (8)</td>
<td>10/11 (3+4)</td>
</tr>
<tr>
<td>Adolescent</td>
<td>8/10 (10)</td>
<td>10/11 (4)</td>
</tr>
<tr>
<td>Adult</td>
<td>8/10 (7)</td>
<td>10/11 (2)</td>
</tr>
</tbody>
</table>

for the models in UML and WUIML for the three member products’ WUIs respectively.

Table 6 shows the ratio of the number of correctly modeled required modeling items via reuse (or product line reuse when WUIML is used) to the total number of reusable required modeling items in modeling the three WUIs and the person-hours spent on each approach.

Table 7 shows the number of correctly modeled required modeling items per person-hours via reuse for the models in UML and WUIML respectively.

The data in Table 3 show that 18 out of 25 required modeling items were correctly modeled for the Pediatric Medical Profile Login WUI using standard UML. On the other hand, the data in Table 3 show that 24 out of 25 required modeling items were modeled for the same Pediatric Medical Profile Login WUI using WUIML. This result in Table 3 indicates that an SME was able to correctly model more required modeling items using WUIML than using standard UML. Table 4 shows the correctly modeled required modeling items and the approximate person-hours for modeling the Pediatric, Adolescent, and Adult Medical Profile Login WUI respectively in standard UML and WUIML. Table 5 shows the calculated value of modeled items per person-hour based on the data from Table 4. The data in Table 5 shows that when using WUIML, an SME was able to correctly model about four times more required modeling items per person-hour (i.e., 9.08) than that in standard UML (i.e., 2.37). Thus, this result indicates WUIML is more efficient than the standard UML in modeling user interfaces for Web-based SPL; therefore, this result supports proposition 1.

Table 6 shows the number of correctly modeled required modeling items via reuse (or product line reuse) and its corresponding person-hours used.
Modeling Variant User Interfaces for Web-Based Software Product Lines

For example, using standard UML to model the Pediatric Medical Profile Login WUI, a SME was not able to reuse any model items out of the ten reusable items in the WUI model. For approximate eight person-hours, the SME has to model without reuse. This is because the Pediatric Medical Profile Login WUI was the first WUI for an SME to model. There is nothing to reuse yet. Notice that the number of reusable required modeling items is only a sub-set of the total number of required modeling items in a WUI model because some items are unique to a particular WUI and cannot be reused.

On the other hand, using WUIML, across all three WUIs (Pediatric Medical Profile Login, Adolescent Medical Profile Login, and Adult Medical Profile Login.), as shown in Table 6, number of correctly modeled required modeling items via reuse are consistent. This is because in WUIML, an SME is to first develop the product line WUI model. The WUI model for the SPL encapsulated all model items for its member product WUI models. Therefore, reusing the items in the product line WUI model creates the WUI model for a member product. As shown in Figure 7, the average number of correctly modeled required modeling items via reuse per person-hours using WUIML (i.e., 2.98) is about four times to that when using standard UML (i.e., 0.65). This result indicates that WUIML enables higher level of reuse than standard UML, thus it indicates increased effectiveness when using WUIML than using standard UML. Therefore, this result supports proposition 2.

The results of the second case study are not shown due to the limitation of the space in the paper. The second case study was conducted by another two SMEs. The units of analysis for the second case study were the software requirements for the Pediatric Exercise Record WUI, the Adolescent Exercise Record WUI, and the Adult Exercise Record WUI. The results from the second case study show that using WUIML to model the Exercise Record WUIs for the three member products of the SPL increases the modeler’s efficiency and effectiveness in modeling. This result is similar to the result from the first case study. The results of the two case studies literally replicate. Based on the results, the analytical generalization that WUIML increases modelers’ efficiency and effectiveness in model WUIs for SPL is drawn in this study.

The generalization thus provides a positive answer to the study’s research question that Web user interface modeling for a Web-based medical SPL using WUIML increases SMEs’ modeling efficiency and effectiveness (thus decreases work).

CONCLUSION AND FUTURE WORK

In standard UML, there is no standard way to model user interface and no standard way to model variant features of user interfaces for a SPL. WUIML defines elements to propose a standard way to model user interfaces as well as
Modeling Variant User Interfaces for Web-Based Software Product Lines

variant user interface features of Web-based SPLs. WUIML improves UML in terms of modeling of user interfaces and use interfaces for SPL.

Case study research method has been applied to investigate whether WUIML method increases modelers’ efficiency and effectiveness on modeling a Web-based medical SPL when compared to standard UML method. The results indicate that modeling WUIs for SPL using WUIML is more effective and efficient than using standard UML.

There are many ways to model software using UML; a human is needed to make the “best” judgment on the correctness of a UML model. In this empirical study, the way to determine whether a required modeling item is modeled correctly or not relies on a human’s modeling experience. Since the same human is using his same modeling experience to make the judgment on all models, the bias due to human judgment should be insignificant.

Since the SMEs are of similar backgrounds and technical experience, the impact on modeling efficiency and effectiveness due to individual SME’s capability should be insignificant. Since the same requirements are given for both approaches (WUIML and standard UML), the reuse opportunity base on application logics (according to the requirements) should be the same for all SMEs.

Future work is to develop a tool to support using WUIML to model WUIs for SPLs and to extend WUIML beyond the Web platform.

REFERENCES

xhtml-modularization.html


*This work was previously published in the International Journal of Information Technology and Web Engineering, Vol. 1, Issue 1, edited by G. Alkhatib and D. Rine, pp. 1-34, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 2.25

A User–Centered Approach to the Retrieval of Information in an Adaptive Web Site

Cristina Gena
Università di Torino, Italy

Liliana Ardissono
Università di Torino, Italy

ABSTRACT

This chapter describes the user-centered design approach we adopted in the development and evaluation of an adaptive Web site. The development of usable Web sites, offering easy and efficient services to heterogeneous users, is a hot topic and a challenging issue for adaptive hypermedia and human-computer interaction. User-centered design promises to facilitate this task by guiding system designers in making decisions, which take the user’s needs in serious account. Within a recent project funded by the Italian Public Administration, we developed a prototype information system supporting the online search of data about water resources. As the system was targeted to different types of users, including generic citizens and specialized technicians, we adopted a user-centered approach to identify their information needs and interaction requirements. Moreover, we applied query analysis techniques to identify further information needs and speed up the data retrieval activity. In this chapter, we describe the requirements analysis, the system design, and its evaluation.

INTRODUCTION

The development of a Web-based information system targeted to different types of users challenges the Web designer because heterogeneous requirements, information needs, and operation modes have to be considered. As pointed out by Nielsen (1999) and Norman and Draper (1986), the user’s mental model and expectations have to be seriously taken into account to prevent her/him from being frustrated and rejecting the
services offered by a Web site. Indeed, this issue is particularly relevant to Web sites offering task-oriented services, because most target users utilize them out of their leisure time, if not at work. Being under pressure, these users demand ease of use as well as efficient support to the execution of activities.

The positive aspect of a technical Web site is, however, the fact that the users can be precisely identified and modeled; moreover, their information needs, representing strong requirements, can be elicited by means of a suitable domain analysis. Therefore, utilities, such as data search and retrieval, can be developed to comply with different goals and backgrounds. Of course, users’ involvement and testing have to be carried out also in this case because they support the development of effective and usable services (see Dumas & Redish, 1999; Keppel, 1991).

In our recent work, we faced these issues in the development of ACQUA, a prototype Web-based information system for the Italian Public Administration presenting information about water resources (a demo is available at http://acqua.di.unito.it). During the system design phase, we put in practice traditional usability principles and adaptive hypermedia best practices and we derived general guidelines for the development of usable Web-based systems for technical users (see Brusilovsky, 1996, 2001; Fink, Kobsa, & Nill, 1999; Maybury & Brusilovsky, 2002). The system described in the rest of this chapter is targeted to two main classes of users:

• Generic users, such as the citizen, who want to be informed about the general health state of rivers, lakes, and underground waters.
• Technical users, such as the public administration employees, who retrieve specific pieces of information for analysis purposes.

In this chapter we describe the requirements analysis, the design, and the evaluation of ACQUA, focusing on the user-centered approach adopted in the prototype design and development phases. We involved domain experts and end users since the beginning of our work in order to assess the usefulness and suitability of the functionality offered by the system, as well as of its user interface. For further information about the system, see Gena and Ardissono (2004).

The rest of this chapter is organized as follows: Section “Background” provides an overview of the relevant user-centered design research. Section “The ACQUA Project” presents our work. Specifically, Section “Application Requirements” describes the interaction and user interface requirements that emerged during the design phase; Section “Adaptive Features” presents the adaptive features we developed for our system; Section “Association Rules” describes the techniques supporting the personalized information search; Section “Evaluation of ACQUA” presents the results of an evaluation we carried out to test the system functionality with real users; and Section “Comparison with Other Solutions” compares our proposal with some related work. Finally, section “Future Trends” discusses some open technical issues and suggests how to address them, and Section “Conclusion” concludes the chapter.

BACKGROUND

Several researchers suggested to address usability issues by developing adaptive systems. For instance, Benyon (1993) proposed adaptivity as a solution, because a single interface cannot be designed to meet the usability requirements of all the groups of users of a system. However, it is possible to prove that adaptivity enhances the usability of a system only if it can be shown that, without the adaptive capability, the system performs less effectively. Benyon identifies five interdependent activities to be considered when designing an adaptive system:
A User-Centered Approach to the Retrieval of Information in an Adaptive Web Site

1. **Functional analysis**, aimed at defining the main functions of the system.
2. **Data analysis**, concerned with understanding and representing the meaning and structure of data in the application domain.
3. **Task knowledge analysis**, focused on the cognitive characteristics required by the system users, such as the user’s mental model, cognitive load, and the required search strategy.
4. **User analysis**, aimed at determining the scope of the user population to which the system is targeted. This analysis concerns the identification of the user attributes that are relevant for the application, such as required intellectual capability, cognitive processing ability, and similar. The target population is analyzed and classified according to the aspects of the application derived from the above-mentioned points.
5. **Environment analysis** is aimed at identifying the characteristics of the environment in which the system is going to operate.

Notice that these phases are similar to the steps followed during the requirements analysis phase of a generic software system (Preece, Rogers, Sharp, & Benyon, 1994). Benyon underlined the fact that adaptive systems should benefit more than other systems from a requirements analysis before starting any kind of evaluation, because the development of these systems has to take a high number of features into account. The recognition that an adaptive capability may be desirable leads to an improved system analysis and design. As a demonstration, he reported an example of an adaptive system development, wherein he prototyped and evaluated the system with a number of users. Several user characteristics were examined to determine their effects on the interaction. Then, further task knowledge and functional analysis were carried out.

Also Oppermann (1994) proposed a user-centered perspective and suggested a design-evaluation-redesign approach. He noticed that the adaptive features can be considered as the main part of a system and thus have to be evaluated during every development phase. The problem is circular:

- A problem solvable by means of the adaptivity has to be identified.
- The user characteristics related to the problem have to be selected.
- Ways of inferring user characteristics from interaction behavior have to be found.
- Adaptation techniques offering the right adaptive behavior have to be designed.

This process requires a bootstrapping method: first some initial adaptive behavior is implemented, then tested with users, revised, and tested again. The reason is that it is hard to decide which particular adaptations should be associated to specific user actions. Furthermore, the adaptations must be potentially useful to the user. The necessity of an iterative process is due to the fact that the real behavior of users in a given situation is hard to foresee; therefore, some evidence can be shown only by monitoring the users’ activity. From the iterative evaluation point of view, the design phases and their evaluation have to be repeated until good results are reached.

Oppermann’s iterative process is very similar to the user-centered system design approach originally phrased by Gould and Lewis (1983) and extended by Norman and Draper (1986).

Dix, Finlay, Abowd, and Beale (1998) pointed out that the iterative design is also a way to overcome the inherent problems of incomplete requirements specification, as only a subset of the requirements for an interactive system can be determined from the start of the project. The iterative evaluation process requires empirical knowledge about the users’ behavior from the first development phases. In the case of an adaptive system, prior knowledge about the real users, the context of use, and domain experts facilitates the
A User-Centered Approach to the Retrieval of Information in an Adaptive Web Site

selection of the relevant data for the user model, such as personal features, goals, plans, domain knowledge, and context. Deep knowledge about users also offers a broad view of the application goals and prevents the system designer from serious mistakes, especially when dealing with innovative applications.

Petrelli, De Angeli, and Convertino (1999) proposed the user-centered approach to user modeling as a way to move from designer questions to guidelines by making the best use of empirical data; they advocated incremental system design as a way to satisfy large sets of users. They reported that at the early stage of development of a mobile device presenting contextual information to museum visitors, they decided to revise some of their initial assumptions about the user model. Indeed, they made this decision after having analyzed the results of a questionnaire distributed to 250 visitors. For instance, they discarded the former user modeling techniques based on stereotypes (because the sociodemographic and personal data taken in consideration did not characterize the users’ behavior in a satisfactory way) in favor of a socially oriented and context-aware perspective. For instance, they noticed that people do not like to visit museums on their own and prefer looking at paintings to interacting with a device.

As discussed by Höök (2000), intelligent user interfaces may violate many usability principles developed for direct manipulation systems. The main problem is that these systems may violate many good principles, such as enabling the user to control the system, making the system predictable (given a certain input, the system always generates the same response), and making the system transparent so that the user understands at least partially how it works. In addition, most adaptive interface developers are more concerned with defining inference strategies than with interface design. For Höök, intelligent user interfaces sometimes require a new way of addressing usability, different from the principles outlined for direct-manipulation systems. Instead of measuring factors such as task completion time, number of errors, or number of revisited nodes, other aspects have to be considered. For instance, “if the system should do information filtering, then we must check whether subjects find the most relevant information with the adaptive system and not necessarily whether they find it fast. This is not to say that the traditional measurements are always wrong—this of course depends upon the task that user and (adaptive) system should solve together” (Höök, 2002, p. 12).

Finally, Palmquist and Kim (2000) investigated the effects of (field independent and field dependent) cognitive style and online database search experience on WWW search performance. They concluded that cognitive style significantly influences the performance of novice searchers. In contrast, experienced searchers display a common behavior: they usually do not get lost in Web pages including many links, but they are able to choose useful navigation strategies. Therefore, Palmquist and Kim suggested that novice users should benefit from Web pages that have a simple design and few links providing information necessary to perform analytic search.

THE ACQUA PROJECT

Application Requirements

In 2003, the Water Resources Division (Direzione Risonse Idriche) of the Piedmont Region and the University of Torino started a project for the development of ACQUA, a Web-based information system presenting data about water resources derived from the monitoring activities on the territory. The goal was to make information available on a Web site that describes the Division and supports a search for data in real time, in order to limit the distribution of information on a one-to-one basis via e-mail messages and paper publications. The technicians of the Division guided us in the system development
by specifying application requirements and by sharing with us a repository of e-mail messages they exchanged with users asking for information throughout the years. The repository provided us with evidence about the users’ interest in water resources data, the inspected information, and the regularities in the search for data. Most questions were posed by the following:

- Employees of other Public Administrations, such as technicians and researchers, who are often interested in environmental impact studies, construction feasibility studies, and historical data.
- Technicians, such as companies working at the construction of bridges and houses.
- Attorneys, who are typically interested in the examination of data concerning specific regions, for example, as a consequence of an environmental disaster.
- Farmers, who wish to monitor the biochemical state of their fields.
- Students attending secondary school, university, and also doctoral programs. These users collect information for the preparation of reports concerning, for example, historical changes in biological and chemical composition of waters, or the evolution of the capacity and hydrometric levels of rivers, and similar.

Following a user-centered approach, we developed the system by involving domain experts and end users since the first design phases. After a requirements analysis phase, we developed a number of mock-ups, which we discussed and redesigned after several focus group sessions with the experts and the users involved in the project. We decided to adopt a cooperative design approach (Greenbaum & Kyng, 1991) in order to utilize the experience of domain experts and technicians in the design of an effective user interface. We based the development of our first prototype on the collected feedback. As the ACQUA system is devoted to the Public Administration, we had to satisfy usability and predictability requirements that imposed the design of a simple user interface. Specifically, our interlocutors suggested the following:

- The interface should be usable and intuitive in order to satisfy user needs and expectations. This first requirement should be followed in every interface design project; however, Public Administrations have the mandatory goal of satisfying all the citizens, thus usability is also intended as a service for the collectivity.
- The system behavior should be highly predictable (Dix et al., 1998) to support first-time visitors in their search task, but also to avoid frustrating professional users who would regularly use it at work. Notice that the predictability requirement has some subtle aspects: for instance, not only the user should foresee what is going to happen next, but also what should not be expected from the service. This is very important to prevent the user from starting the exploration of paths that will not provide her/him with the information, or the functionality (s) he is looking for.
- The system should provide the user with data that can be analyzed without preprocessing. Therefore, search results should be presented in machine-processable formats, in addition to the pictorial ones suitable for a general-purpose presentation in Web pages.
- For the sake of accessibility, the pages of the user interface should be optimized for standard browsers, without the need of special equipments or software environments.

In order to maximize the usefulness of the information that can be retrieved from the Web site, we decided to make the system generate the search results in formats, such as MS Excel® tables and textual (TXT) files, directly supporting the
data analysis and interpretation at the user side. We also tried to address efficiency in the retrieval of information by reconsidering the design of the general Web site to be presented. We wanted to offer the right information the user is looking for; thus, we decided to show the main search functions in the home page of the Web site, and to move textual information, such as the pages describing the Public Administration divisions, in secondary pages, which can be reached by following hypertextual links.

Moreover, having analyzed the data about the users interested in water resources, we identified two main targets to which the system should adapt. For shortness, we denote these categories as novices and experts.

- **Novice users**, such as students and generic citizens, visit the Web site on an occasional basis and are not familiar with the content presented by the information system.
- **Expert users**, such as technicians, farmers, and the personnel of other Public Administrations, frequently visit the site and are familiar with the domain-specific information provided by the system.

In order to take the interaction requirements of these users into account, we defined two search functions:

i. The **simple search** is a geographical search modality and guides the user step by step in the retrieval of information;

ii. The **advanced search** offers forms where the expert user may compose the queries in single step. Figure 1 shows the user interface of the ACQUA prototype supporting

---

**Figure 1.** Searching quantitative data (continuous hydrometric and chemical-physical parameters) about Po River in the Torino-Murazzi observation point
the advanced search; the menus enable the user to specify the river (“Scegli il corso d’acqua”), observation point (“Scegli il punto di monitoraggio”), start date (“Data Inizio”), and end date (“Data Fine”). Moreover, the user interface enables the user to select the hydrometric and chemical-physical parameters to be inspected.

Thus, novice users may search for information in a friendly modality and the eligible choices are restricted and presented along the path, while expert users benefit from a faster search function.

As a matter of fact, the information about water resources exploited by the system is unavoidably incomplete. For instance, some data are collected by automatic stations, which have been set up at different times over the years and sometimes are out of order. Moreover, unfortunately, data collected in manual observation points have been stored in unstructured formats and the historical series has been reconstructed only for the very recent past.

For the sake of predictability, the simple and advanced search functions prevent the user from composing any queries that are incorrect, or are aimed at searching for unavailable data. The idea is that, in both cases, the system should only present the choices leading to available results. For instance, as shown in Figure 1, the labels of the parameters, which are not available for the Po River, are shaded and cannot be selected by the user to define a query.

**Adaptive Features**

The information about water resources concerns rivers, lakes, and underground waters and includes the following:

- Descriptive data about resources and observation points: for example, maps of the points, charts representing environmental changes, pictures, documents, publications, and descriptions of the monitoring stations. For instance, Figure 2 (“Caratteristiche della stazione di monitoraggio TORINO” [“Features of the Torino monitoring station”]) shows the coordinates and other information about the observation point on Po River located in Torino, Parco Michelotti.

- Measurement parameters concerning physical dimensions and other features, which characterize the environmental state of the resources. These parameters are grouped in two main classes:
  - **Qualitative parameters**, which are periodically measured: technicians visit the observation points, collect data, and take samples for laboratory tests.
  - **Quantitative parameters**, which are monitored by automatic stations.

These stations carry out the measurements on a daily basis.

The ACQUA Web site is organized in four main sections, respectively devoted to the presentation of qualitative and quantitative information about rivers, information about lakes, and information about underground waters. The system enables the user to retrieve data about water resources by performing a simple or advanced search in all the sections of the site. Therefore, a large amount of heterogeneous data is accessible, ranging from biological and chemical data to capacity measurement and hydrometric levels (for details, see Gena & Ardissono, 2004).

We noticed that, by performing queries aimed at selecting a large number of data items, belonging to different categories, the results returned by the system were complex and hard to present in an intuitive results table. However, as shown by the repository of user requests we analyzed, users often need to combine heterogeneous data to accomplish their goals. For example, in construction feasibility studies, users are interested in qualitative and quantitative parameters of
A User-Centered Approach to the Retrieval of Information in an Adaptive Web Site

rivers and underground waters, considering the historical series. In order to keep the user interface simple and to guarantee that the presented results are not confusing, we decided to limit the user’s freedom in composing the queries: to retrieve very heterogeneous types of information, the user must define more than one search query. For example, as shown in Figure 1, the ACQUA query interface enables the user to choose from different rivers, observation points, years, and data types. Other categories, such as qualitative and quantitative data about rivers, lakes, and underground waters are treated as separate sections of the Web site and have their own query functions.

Unfortunately, although this approach enforces the clarity of the results, it makes the search for multiple types of information a lengthy task. Therefore, a compromise between clarity and efficiency must be found. In order to address this issue, we extended the system with an intelligent search component, which complements the user’s explicit queries with follow-up queries (Moore & Mittal, 1996) frequently occurring together in navigation paths. When possible, the system anticipates the user’s queries and makes the extended search results available as personalized suggestions that can be downloaded on demand. If the user is interested in the recommended information, (s)he can retrieve it by clicking on the adaptive suggestion links, without performing any further queries. At the same time, the system retrieves the extended results only after
the user clicks on a suggestion link in order to avoid precaching possibly useless data.

For instance, Figure 3 shows the recommendations generated by the system in the lower portion of the page (“Ti consigliamo anche i valori dei parametri chimici e microbiologici” [“We also suggest results about chemical and microbiological parameters”]).

During different interaction sections, the same user may be interested in rather different types of information; therefore, we decided to base the system’s recommendations on the analysis of her/his navigation behavior, leaving the management of a long-term user model apart. One immediate advantage is the fact that the user can interact with the system in an anonymous way, without signing up for the service. The follow-up queries are generated as follows: the search queries performed by the user while (s)he browses the Web site are monitored and association rules which suggest other, strictly related queries are applied.

Each association rule has a condition part specifying constraints on the previous navigation behavior, and an action part defining a relevant follow-up query to be performed in order to retrieve complementary information. The rules we defined in our current prototype are mutually exclusive and they are selected and fired by applying a very tiny and efficient inference engine. This engine would not be suitable to manage a large set of conflicting rules: a general-purpose rule-based engine should be employed to that purpose. However, we prefer to maintain a simple set of adaptation rules, and to avoid embedding complex rule-based engines in order to keep the adaptive features as lightweight as possible. In fact, the management of the interaction is subject to a relevant overload due to the generation of results in multiple formats, which is a main requirement for the information system. In this situation, minimalist but efficient adaptation to the user is strongly preferred to flexible but complex one.

**Association Rules**

In order to define the association rules to be applied for anticipating the user’s information needs, we analyzed a repository of requests, which real

---

**Figure 3. Annotated link for the suggested information and descriptions of the monitoring stations**
A User-Centered Approach to the Retrieval of Information in an Adaptive Web Site

users posed to the Water Resources Division over the years; the requests consisted of e-mail messages and fax documents. As we noticed that different kinds of information frequently occurred together in these requests, we decided to analyze the frequency of co-occurrence in order to identify the regularities. Specifically, we analyzed 97 requests and we selected a set of features describing the requests in a systematic way. These features concerned rather different aspects of the requests; thus, for clarity purposes, we grouped them in subcategories. In the following, we report the subcategories we defined and for each one we list some sample features:

- **Kind of request:** for example, environmental impact study, construction feasibility studies, and lawyers’ studies.
- **Request features:** for example, information about one or more rivers, about lakes or underground waters, about one or more observation points on a river or lake.
- **Kind of data:** for example, qualitative or quantitative parameters, biological and chemical data, hydrometric level, average daily capacity.
- **Data features:** for example, raw data, or elaborated data such as medium, maximum, and minimum values during a time interval.
- **User features:** for example, research center, Public Administration, technicians, farmers, and attorneys.

We computed the frequency with which the features co-occurred in the requests: if the frequency exceeded a given threshold, the set of involved features became a possible candidate for an association rule. Then we compared the extracted associations with their original requests in order to validate our findings with factual knowledge, and finally we asked the technicians of the Water Resources Division if our conclusions were correct. After this last check, we selected the correct associations and we encoded the rules in the system.

For instance, a rule suggests to retrieve qualitative parameters about a water resource if the user has asked for quantitative historical data for more than one observation point on that resource, supposing that (s)he is looking for information for a construction feasibility study. Another rule suggests retrieving the environmental state indexes of a resource if the user has requested biological and chemical data, under the hypothesis that (s)he is involved in an environmental impact study.

**Evaluation of ACQUA**

We first evaluated the ACQUA prototype in a usability test by involving external users who were not cooperating at the project (see Dumas & Redish, 1999, for methodological details). The evaluation highlighted some usability problems concerning the presentation of basic information, such as the choice between simple and advanced search and the background color of the menus. After having solved those problems, we tested the final prototype with real end users representative of the users the Web site is devoted to. In particular, we involved technicians working at the Water Resources Division in different fields (rivers, lakes, underground rivers, etc.) and not collaborating to the design of the project. We carried out both an experimental evaluation and a qualitative session to assess the suitability of the adaptive features offered the system.

**Subjects.** We evaluated 10 potential users of the ACQUA system, four females and six males, aged 30–50. All the users worked in the water resource area and none of them was involved in the project.

**Procedure.** The subjects were split up in two groups (five subjects each) and randomly assigned to one of the two groups. The experimental group had to solve some tasks using the adaptive Web site, which applies the association rules described in Section “Adaptive Features” to compute the
results of follow-up queries related to the users’ explicit queries. Instead, the control group had to solve the tasks without adaptation.

**Experimental tasks.** Every subject had to solve seven tasks, each one representing a real task the user can perform in the Web site. As suggested by our correlation study, the tasks were strictly correlated and could be grouped in three search activities the user often performs together. The first activity conveyed the whole information useful to an environmental impact study. The second one supported construction feasibility studies. The third activity supported lawyers’ studies and activities.

- In the control group, the users had to submit a new query for every task, in order to obtain the requested results. The new queries were submitted by filling in the query specifica-

- In the experimental group, the users could obtain the extra information related to the next task to be performed by clicking on an adaptive suggestion link that supports the immediate retrieval of the suggested information (see, e.g., Figure 3).

**Experimental design.** Single-factor (the adaptivity) between-subjects design.

**Measures.** The subjects’ navigation behavior was recorded by using Camtasia Studio®. We measured the task completion time and then the subjects’ satisfaction, by means of a post-task walk-through.

**Hypothesis.** We hypothesized that the users working in the experimental group could obtain better performance results than those of the control group.

**Results.** The ANOVA (analysis of variance) showed that the subjects of the experimental group achieved the best performance results. In addition, we calculated the effect size (treatment magnitude) and the power (sensitivity) as suggested in Chin (2001). The effect size ($\omega^2$) measures the strength, or the magnitude, of the treatment effects in an experiment. In behavioral sciences, small, medium, and large effects of $\omega^2$ are 0.01, 0.06, and $>0.15$, respectively. The power of an experiment ($n'$) is the ability to recognize treatment effects and the power can be used for estimating the sample size. In social science, the accepted value of the power is equal to 0.80, which means that the 80% of repeated experiments will give the same results. In the following, we show a summary of the results:

**Task 2.**

ANOVA: $F(1.8) = 12.45$  \( p<0.01; \)
$\omega^2=0.53; \ n'=3.49$

**Task 3.**

ANOVA: $F(1.8) = 12.12$  \( p<0.01; \)
$\omega^2=0.53; \ n'=3.60$

**Task 5.**

ANOVA: $F(1.8) = 14.16$  \( p<0.01; \)
$\omega^2=0.57; \ n'=3.04$

**Task 7.**

ANOVA: $F(1.8) = 9.23$  \( p<0.05; \)
$\omega^2=0.45; \ n'=4.86$

It should be noticed that all the results are significant and have a large estimate of the magnitude of the treatment effect. In addition, by exploiting a power of 0.80 and the corresponding $\omega^2$ for each task we could determine the requested sampled size, which fits our sample size ($n=5$) (for details about statistics, see Keppel, 1991).

**Post-task walk-through.** During any post-task walk-through, test subjects are asked to think about the event and comment on their actions. Thus, after each test we talked to the subjects to collect their impression and to discuss their performance and the problems encountered during the test. In this session, we also aimed at retrieving useful feedback for a qualitative evaluation of the site. In fact, although our experimental evaluation
reported significant results supporting our hypothesis, the actual user behavior could be different. As recently pointed out by Nielsen (2004), statistical analyses are often false, misleading, and narrow; in contrast, insights and qualitative studies are not affected by these problems because they strictly rely to the users’ observed behavior and reactions.

In most cases, the interviewed users were satisfied with the site. Most of them encountered some problems in the execution of the starting query of task 2, thus we modified the interface form.

- All the users of the experimental group followed the adaptive suggestion link provided by the system but they did not realize that it represented a personalization feature. When we explained the adaptations, they noticed the particularity of the suggestion (“We also recommend you ...”). Anyway, they were attracted from the suggestions and they appreciated the possibility of skipping the execution of a new query. The adaptive suggestions were considered visible and not intrusive.

- The users of the control group reported similar considerations when we described the adaptive features offered by the Web site. Even if they did not receive any suggestions during the execution of tasks, they explored the result pages in order to find a shortcut to proceed in the task execution. After having followed some links, they went back to the previous query page or to the home page by clicking on the “Back” button of the browser.

Both groups displayed a common behavior pattern: the users explored a results page before starting a new search. Nevertheless, their behavior could be influenced by the test condition, because tested users tend to pay a lot of attention to their own actions and to the page design.

We conclude by admitting that although the test subjects were satisfied with the adaptation features, only the real system usage can demonstrate our hypothesis. However, both quantitative and qualitative test results are encouraging and we think that the adaptations are correctly placed. After this test, we presented the adaptive version of the Web site to the technicians of the Water Resources Division collaborating on the project. They confirmed the correctness of association rules we defined and they decided to replace the non-adaptive version of the prototype system with the adaptive one.

**Comparison with Other Solutions**

The ACQUA system has a plain user interface, designed to meet simplicity, usability, and predictability requirements, but it offers advanced interactive features enabling the user to create a personal view of the information space. Two search features, targeted to novice and expert users, are available, and the search results are presented in both pictorial and machine-readable formats in order to support direct data manipulation at the userside. Moreover, the system analyzes the user’s queries to identify her/his information needs, and it employs association rules to propose follow-up queries complementing the search results with strictly related information. The follow-up queries are applied on demand; thus, the user can ignore them if (s)he is not interested in the additional data, and the system does not need to retrieve any uninteresting information.

The advanced search features we presented differ from the related work in various aspects. On the one hand, the inferences performed by our system are simpler than the probabilistic ones applied in other automated assistants, such as Lumiére (Horvitz, Breese, Heckerman, Hovel, & Rommelse, 1998) and ACE (Bunt & Conati, 2003), which exploit Bayesian networks to capture the dependencies among the user actions.
The point is that the user interacting with the ACQUA system does not carry out a complex task requiring a problem-solving activity. Therefore, lightweight rules associating contextually related search queries are sufficient to predict the implicit information needs and to complement the search for information accordingly. Our approach also differs from the follow-up question answering techniques proposed by Moore and Mittal (1996): in order to efficiently manage the query selection process, our follow-up queries are precompiled in a set of association rules, instead of being generated by a planner.

On the other hand, we apply query analysis techniques to identify regularities in search patterns. This differs from the typical inferences carried out in recommender systems, which reason about the features of the selected items to identify the user’s priorities (see, e.g., Billsus & Pazzani, 1999), or about the regularities in the selection of individual items (see, e.g., the work by Cotter & Smyth, 2000; GroupLens, 2002).

Liu, Yu, and Meng (2002) propose other query analysis strategies for personalized Web search. However, instead of personalizing the proposed results, their system supplies a small set of categories as a context for each query. The system combines the user’s search history with a general user profile automatically extracted from a category hierarchy to offer a personalized context for disambiguating the proposed query results. In ACQUA, we do not manage long-term user preferences because we noticed that, in different interaction sections, the same users are interested in rather different types of information. We thus decided to base the recommendations only on the analysis of the user’s search behavior.

**FUTURE TRENDS**

It is worth mentioning that the manual definition of the first set of association rules supporting the user’s search task was a lengthy work and might not be easily replicated to revise the rules along time. However, if the Water Resources Division employs the ACQUA system as its official Web site, the log files generated by the system will provide structured evidence about user behavior (in addition to e-mails and faxes). Thus, data-mining techniques could be exploited to automatically recognize usage patterns and revise the association rules accordingly.

Indeed, we believe that these techniques can support the analysis of user behavior in an effective way, but they still have to be coupled with human analysis, in order to validate and interpret results: in several cases, these techniques have generated some very interesting results, but also other irrelevant or hardly understandable findings, which have been discarded.

At any rate, Web usage mining techniques, derived from machine learning methods such as knowledge discovery in data (KDD or data mining) can contribute to automate the adaptation of Web-based systems to the users. According to the scheme proposed by Pierrakos, Paliouras, Papatheodorou, and Spyropoulos (2003), ACQUA can be classified as a Web personalization system offering task performance support: this functionality involves the execution of a particular action on behalf of the user. In our case, the system generates queries and makes the results available as links to some files storing them. This functionality is considered as the most advanced personalization function and it is seldom offered by Web-based personalized services.

The most suitable data-mining technique, given the adaptive goals of the ACQUA system, is the **sequential pattern discovery**, which is aimed at identifying navigational patterns (event sequences) in the analyzed data (in our case, Web usage data). This methodology supports the discovery of event sequences that can be summarized as follows: “If event A, B, and C occur in that order, then events D, E, and F always follow.”
Two types of methods are generally applied to discover sequential patterns: deterministic techniques, which record the navigational behavior of the users and extract knowledge from the analyzed data, and stochastic methods, which use the sequence of already-visited Web pages to predict the behavior occurring in the next visits. Once sequential patterns have been discovered, the extracted knowledge can be automatically integrated in the personalization process, and the system behavior adapted accordingly.

CONCLUSION

We presented our experience in the design and development of ACQUA, an interactive prototype Web site for the Public Administration. The system presents information about water resources and supports the user in the search for generic information, as well as technical information about the rivers, lakes, and underground waters.

The usability and functional requirements that emerged during the design of the ACQUA system were very interesting and challenging, as they imposed the development of functions supporting the efficient retrieval of data by means of a simple user interface. We found out that the introduction of basic adaptivity features, aimed at understanding the user's information needs in detail, was very helpful to meet these requirements.

We were asked to develop a system having a simple user interface, designed to meet usability and predictability requirements. This fact limited our freedom to add advanced interaction features, desirable in a Web site visited by heterogeneous users; however, it challenged us to find a compromise between functionality and simplicity. In order to address this issue, we developed two interactive features enabling the user to create a personal view on the information space:

- The system offers a simple and an advanced search functions targeted to novice and expert users, respectively.
- Moreover, the system carries out a query analysis aimed at identifying the user's information needs, and applies association rules to extend the user's queries and complete the search results with data that is usually retrieved together by end users.

Qualitative and quantitative evaluation results showed that the adaptive user interface was more successful than the nonadaptive one. The reason was probably the concrete help offered by the adaptive suggestions, which speed up the execution of time-consuming search tasks. Moreover, the adaptive features were not perceived as intrusive and the user was allowed to skip useless suggestions. Furthermore, the system did not impose a previous annoying and discouraging registration phase.

As discussed in Section “Future trends,” the adaptive features offered by the ACQUA system could be improved by the integration of Web-usage mining techniques aimed at discovering real usage patterns. In that way, the association rules employed to identify the user's implicit information needs could be automatically updated along time. However, we believe that the rules we manually defined provide a knowledge base that cannot be replaced with automatically extracted rules. In principle, both kinds of rules could be integrated in order to enhance the effectiveness of the system adaptations.

ACKNOWLEDGMENTS

This work was funded by Regione Piemonte, Direzione Risorse Idriche. We thank Giovanni Negro, Giuseppe Amadore, Silvia Grisello, Alessia Giannetta, Maria Governa, Ezio Quinto, Matteo
Demeo, and Vincenzo Pellegrino, who assisted us during the system development and provided the domain-specific knowledge.

REFERENCES


A User-Centered Approach to the Retrieval of Information in an Adaptive Web Site


This work was previously published in Cognitively Informed Systems: Utilizing Practical Approaches to Enrich Information Presentation and Transfer, edited by E. M. Alkhafia, pp. 142-166, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 2.26
Auto-Personalization Web Pages

Jon T. S. Quah
Nanyang Technological University, Singapore

Winnie C. H. Leow
Singapore Polytechnic, Singapore

K. L. Yong
Nanyang Technological University, Singapore

INTRODUCTION

This project experiments with the designing of a Web site that has the self-adaptive feature of generating and adapting the site contents dynamically to match visitors’ tastes based on their activities on the site. No explicit inputs are required from visitors. Instead a visitor’s clickstream on the site will be implicitly monitored, logged, and analyzed. Based on the information gathered, the Web site would then generate Web contents that contain items that have certain relatedness to items that were previously browsed by the visitor. The relatedness rules will have multidimensional aspects in order to produce cross-mapping between items.

The Internet has become a place where a vast amount of information can be deposited and also retrieved by hundreds of millions of people scattered around the globe. With such an ability to reach out to this large pool of people, we have seen the expulsion of companies plunging into conducting business over the Internet (e-commerce). This has made the competition for consumers’ dollars fiercely stiff. It is now insufficient to just place information of products onto the Internet and expect customers to browse through the Web pages. Instead, e-commerce Web site designing is undergoing a significant revolution. It has become an important strategy to design Web sites that are able to generate contents that are matched to the customer’s taste or preference. In fact a survey done in 1998 (GVU, 1998) shows that around 23% of online shoppers actually reported a dissatisfying experience with Web sites that are confusing or disorganized. Personalization features on the
Web would likely reverse this dissatisfaction and increase the likelihood of attracting and retaining visitors.

Having personalization or an adaptive site can bring the following benefits:

1. Attract and maintain visitors with adaptive contents that are tailored to their taste.
2. Target Web contents correspondingly to their respective audience, thus reducing information that is of no interest to the audience.
3. Advertise and promote products through marketing campaigns targeting the correct audience.
4. Enable the site to intelligently direct information to a selective or respective audience.

Currently, most Web personalization or adaptive features employ data mining or collaborative filtering techniques (Herlocker, Konstan, Borchers, & Riedl, 1999; Mobasher, Cooley, & Srivastava, 1999; Mobasher, Jain, Han, & Srivastava, 1997; Spiliopoulou, Faulstich, & Winkler, 1999) which often use past historical (static) data (e.g., previous purchases or server logs). The deployment of data mining often involves significant resources (large storage space and computing power) and complicated rules or algorithms. A vast amount of data is required in order to be able to form recommendations that made sense and are meaningful in general (Claypool et al., 1999; Basu, Hirsh, & Cohen, 1998).

While the main idea of Web personalization is to increase the ‘stickiness’ of a portal, with the proven presumption that the number of times a shopper returns to a shop has a direct relationship to the likelihood of resulting in business transactions, the method of achieving the goal varies. The methods range from user clustering and time framed navigation sessions analysis (Kim et al., 2005; Wang & Shao, 2004), analyzing relationship between customers and products (Wang, Chuang, Hsu, & Keh, 2004), performing collaborative filtering and data mining on transaction data (Cho & Kim, 2002, 2004; Uchyigit & Clark, 2002; Jung, Jung, & Lee, 2003), deploying statistical methods for finding relationships (Kim & Yum, 2005), and performing recommendations bases on similarity with known user groups (Yu, Liu, & Li, 2005), to tracking shopping behavior over time as well as over the taxonomy of products. Our implementation works on the premise that each user has his own preferences and needs, and these interests drift over time (Cho, Cho, & Kim, 2005). Therefore, besides identifying users’ needs, the system should also be sensitive to changes in tastes. Finally, a truly useful system should not only be recommending items in which a user had shown interest, but also related items that may be of relevance to the user (e.g., buying a pet => recommend some suitable pet foods for the pet, as well as suggesting some accessories that may be useful, such as fur brush, nail clipper, etc.). In this aspect, we borrow the concept of ‘category management’ use in the retailing industry to perform classification as well as linking the categories using shared characteristics. These linkages provide the bridge for cross-category recommendations.

**DESCRIPTION OF SYSTEM**

In this article, we seek to provide an adaptive feature using a fast and cost-effective means. The aim is to provide adaptiveness in the sense that when a visitor selects the next link or a new page, the contents of the page generated will have relatedness to previous pages’ contents. This adaptive feature will be immediate and will not experience delay or repetitive computational filtering problems, as compared to using mining or collaborative filtering (Claypool et al., 1999; Basu et al., 1998).

The rules-based technique offers an excellent and flexible mechanism to specify rules that map categories that exhibit relatedness among themselves (IBM, 2000). Adding new product
lines is simple, by just adding new sets of rules to map the new products accordingly. For direct item-to-item relatedness mapping, it is not so scalable and feasible to implement through use of the rules-based technique. Instead we will use content-based filtering for generating direct item-to-item mappings. The content-based technique (Claypool et al., 1999; Basu et al., 1998) allows item-to-item mapping to be implemented in a scalable manner by just defining the item’s attribute, and the recommendation engine will automatically generate or match items of same attribute without involving user efforts (Basu et al., 1998).

In order to facilitate the deployment of these recommendation techniques, the Web domain is structured into their respective categories that exhibit relatedness among them. For example, pet dog would have relatedness to books on dogs. Each of the categories is given a unique ID value. The relatedness rules make use of these IDs to generate recommendations. The Web site domain is structured into supernodes (SA, SB...) which branch into child nodes (A1,A2...An;...K1,K2...Kn). These supernodes are a representation of products on the Web site, and the child nodes are used to represent the breakdown of the products into categories. Below each of the child nodes are the sub-child nodes (Aa1,...Axm;...Ka1,Ka2,...Kxm) that represent the items. Each of the child nodes (A1,A2...An;...K1,K2...Kn) is identified with its corresponding ID value. With this structure, rules-based mapping can be easily identified and applied among the child nodes by defining the IDs that will result into a recommended page.

The syntax of a relatedness rule is:

\[ \text{ID}_a:\text{ID}_b:\ldots\Rightarrow \text{Target page} \]

The entries \( \text{ID}_x \) represent the IDs of Web pages that have relatedness and thus can be mapped directly to the Web page (link) identified as target page. A rule is considered matched when any of the Web page IDs in the rule is also found to exist in the selector’s list (visitor’s profile). The selector’s list is compared against all the rules in the rule file. Only one of the rules will be used, and that is the rule that has the most number of \( \text{ID}_x \) elements matching the selector’s list. In the event of tie, the rule which matches with selector entry that carries the higher points will be used, or if they still tie, then precedence of rule entry in the rule file will be used to determine the final rule to use.

Mappings across sub-child nodes are done based on their attributes (content based). A file will be used to define the list of items (or contents) of each Web page (at category level) and also any of attributes for each of the items. The syntax for this item list file is:

\[ \text{Id}_x:\text{Id}_y:\ldots:\text{attr}_a:\text{attr}_b\Rightarrow\text{display.html} \]

The entries \( \text{Id}_x \) are indexes to the html page that contains all the information regarding the item. An example of these indexes could be the type of pet. The information about each type of pet is found in the corresponding html file. An item sorter engine will, at startup, sort out the indexes and their corresponding html files. This will allow the page contents to be arranged or displayed based on any of these indexes. The other set of entries \( \text{attr}_x \) define the attributes of the item. When the visitor selects any of the items for browsing, the corresponding \( \text{attr}_x \) will be tracked. Whenever the visitor browses a new page, the recommendation engine will check for items that have the same \( \text{attr}_x \) and automatically include the corresponding html file for display (see Figure 1).

Our prototype system also incorporates means of generating a dynamic profile that changes as the visitor browses through the Web site. Implicit tracking does not require any explicit inputs or intervention by the visitor. Conventionally this is done either through use of user authorization, hidden form fields, URL rewriting, or cookies (Hunter et al., 2002). Although cookies offer an
elegant solution to the stateless HTTP protocol when compared to the other three, it is also frowned upon by some visitors and as such are disabled (Mobasher et al., 1999; Cooley, Mobasher, & Srivastava, 1999). Instead, the tracking of visitors’ activities is implemented through use of a session object on the server side (Hunter et al., 2002). This object will be used to record all the Web pages (tagged with IDs on the server side) browsed by the visitor during the current session, hence obtaining the profile of the visitor. However, a cookies mechanism is still implemented to provide the “remember” feature, in the sense that if the cookies feature is enabled on the visitor’s browser, the activities on the Web site from the most recent visit will be available (and retrieved) during the next visit. This allows the system to build the visitor’s profile immediately. Hence we are able to combine the static profiling (from cookies) with the dynamic one (tracked by session object) to build a profile that can “remember” a visitor’s previous preferences and dynamically change itself as the visitor continues browsing the Web site.

In order to give priority to activities that occur more often and recent, a point system is used whereby the current session log is given heavier weight than those retrieved from the cookies so that the current activities will be more likely to be nominated into the visitor’s latest profile. The activities log tracked by the session object will be given higher consideration during profiling in order for the system adaptiveness to be reflected accordingly and immediately to the changes in the visitor browsing behavior. In this design, a total of three activities logs are used (two from cookies if available, and the remaining one is from the current session object that tracks the current activities).

In order to demonstrate the adaptiveness feature, the Web site domain should consist items or products that have the following characteristics:

1. Products that can be easily categorized.
2. Selection of items or products should be able to reflect the taste or preference of visitor.
3. Certain form of relatedness between the products.
Auto-Personalization Web Pages

With the above considerations taken into the account, we implemented a Web portal using an online shop named My-Pet that sells the following products:

1. Pets
2. Books on pets
3. Accessories for pets
4. Food for pets

To demonstrate how the system functions, we selected the following list of categories:

Pets: Bird, Dog, Cat, Hamster, Fish
Books: How to care…, Diseases of…, Nutrition for …
Accessories: Pet wears, Toys, Restraints
Foods: Meals, Snacks, Supplements

The My-Pet recommendation engine is based on the relatedness determination among its products. The relatedness determination for My-Pet is defined at three levels:

1. Relatedness between categories across products (e.g., pet dog and dog food and books on dogs);
2. Relatedness between categories in the same product (e.g., books on how to care for dogs and books on diseases of dogs and books on nutrition for healthy dogs); and
3. Relatedness at items level across products (e.g., food and accessory items from the same company/brand).

Items (1) and (2) are achieved based on the rules-base technique, while content-based filtering is used for item (3).

The server software architecture is decomposed into independent modules that have very similar structure. Each product has its own module that is responsible for presenting and generating its dynamic Web page contents. The development platform chosen is Java Server Pages (JSP), which offers object-oriented components deployment. By developing the whole Web program in JSP

Figure 2. Software architecture
and Java, the server program is OS-platform independent.

Each module is made up of the following components:

- JSP Page
- Logs Processor (JavaBean)
- Recommendation Engine (JavaBean)
- Item Sorter (JavaBean)

The JSP page is responsible for delegating the tasks to the respective components, and at the end presenting the results into HTML format for viewing by the visitor.

The Logs Processor component serves to log all the activities by the visitor and process the activities logs before submitting to the Recommendation Engine. It plays the role of profiling the visitor’s activities. The Recommendation Engine is mainly used for determining product-to-product mapping based on the selector’s list (profile) generated by the Logs Processor. The Items Sorter’s main function is to read the list of product items and sort them such that they can be retrieved via means of indexing or key. This component is also used to retrieve an item attribute that is used for Content-Based filtering, based on the input determined by the Logs Processor component (see Figure 2).

IMPACT OF SYSTEM

The prototype system is implemented using a Java Server from JRUN which provides the container for running JSP and JavaBeans classes. The database stores all external data inputs. The Web server PC is installed with JRUN and the Web application. Customers can access the Web site using any Web browser.

To illustrate the effectiveness of this system in providing dynamic content generation, a series of user inputs and the subsequent system responses are tabulated.

From the trial run results obtained, it was observed that the system behaved as expected, providing the adaptive feature of generating Web contents based on users’ activities on the Web site (see Table 1).

<table>
<thead>
<tr>
<th>Browsing Activities</th>
<th>Web Content Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ability to automatically present the choice of viewing items in a page (e.g., what type of pets) based on visitor’s previous preference;</td>
<td></td>
</tr>
<tr>
<td>2. ability to automatically present Web contents that display the category that exhibits relatedness (as defined in Rule Files) to other products’ categories browsed by the visitor previously;</td>
<td></td>
</tr>
<tr>
<td>3. ability to generate an option to view list of items that has relatedness to other items that the visitor has visited or selected; and</td>
<td></td>
</tr>
<tr>
<td>4. ability to generate a recommendation based on cookies retrieved from the visitor’s PC.</td>
<td></td>
</tr>
</tbody>
</table>

CONCLUSION

We have developed a system that can adapt its Web contents based on visitors’ activities on the Web site through combining rule-based with content-based filtering techniques—resulting in an implementation that is both flexible and can rapidly adjust its recommendations. Rule-based structure offers cross-product mapping. Content-based filtering takes the items’ attributes into account when generating recommendations. The system transparently and seamlessly tracks the visitor on the server side and does not require explicit inputs (ratings or purchases or login account) to determine the visitor’s profile dynamically.

My-Pet’s system design utilizes the concept of category management, which is widely practiced in brick-and-mortar shop fronts and maps product
Auto-Personalization Web Pages

Table 1. Results obtained from trial runs

<table>
<thead>
<tr>
<th>Browsing Activities</th>
<th>Web Content Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer initially selects Dog category under the Pet section and chooses from the</td>
<td>The system automatically displays the choices of cats for selection. The cats are displayed based on the degree of friendliness/loyalty to owners.</td>
</tr>
<tr>
<td>choice of displaying dogs based on friendliness/loyalty. The customer then</td>
<td></td>
</tr>
<tr>
<td>chooses Cat category under Pet section.</td>
<td></td>
</tr>
<tr>
<td>In continuation from above, the customer then selects the choice of displaying Cats</td>
<td>The system automatically switches from presenting cats to presenting dogs based on their feeding habits.</td>
</tr>
<tr>
<td>based on feeding habits. User then re-selects the Dog category.</td>
<td></td>
</tr>
<tr>
<td>Customer selects Dog category under the Pet section. Customer then chooses Book</td>
<td>The system automatically presents books that are related to dogs.</td>
</tr>
<tr>
<td>product.</td>
<td></td>
</tr>
<tr>
<td>Customer selects Supplements category under Food section. Visitor then selects Petsfriend (from the list of brands) and clicks the item “Best Choice Dog Cookies”. Customer then chooses Book product.</td>
<td>The system automatically brings the customer to Nutrition category under the Book section and also displays the option for viewing book titles related to dog nutrition from Petsfriend.</td>
</tr>
<tr>
<td>Customer selects Dog category under Accessories product and then chooses Dog</td>
<td>The system automatically forwards the visitor to the Dog category when the Food product was selected in this new browsing session.</td>
</tr>
<tr>
<td>category under Books as well. Customer subsequently closes the browser before</td>
<td></td>
</tr>
<tr>
<td>restarting again to browse the Web site again, but this time around selecting the</td>
<td></td>
</tr>
<tr>
<td>Food product instead.</td>
<td></td>
</tr>
</tbody>
</table>

taxonomy into a cyberspace virtual environment. The key to using category management is that it makes system expansion systematic and easy. In our illustration, we included a handful of products and categories. However, the system architecture provides a structure that allows more products or categories to be added easily. Sub-categories can also be added as long as each of them is given a unique ID page. A new set of relatedness rules can then be defined for these newly added elements. This is in fact a strength of rule-based implementation, where new linkages between product categories and subcategories can be added and removed as need arises. Rules may also carry different precedence values.

Such a highly adaptive system does have a weakness—it has taken away some navigation control from the user. It was noted in a 1998 survey (GVU, 1998) that 17% of Web users experience difficulty in returning to pages visited before. A system with an auto-recommendation feature is likely to fair worse in this aspect. Therefore, for
future improvement to the system, a feature could be added to the system whereby a link page is created and this page contains all links to pages that were previously browsed by the user.

In addition, a search option can also be included. This search option indirectly provides a way for the visitor to tell the system his/her taste or preference. This serves as an excellent and accurate input in building a profile that even more closely reflects the visitor’s preferences. Hence the system will be able to use a more accurate visitor’s profile when generating recommendations.

REFERENCES


**KEY TERMS**

**Category Management:** Classify and manage items based on some predetermined categories.

**Clickstream:** A sequence of mouse clicks.

**Collaborative Filtering:** Unveiling general patterns through “sniffing” through user’s past activities.

**Personalization:** Customization to individual user’s preferences and needs.

**Portal:** The entry node/point of navigation semantic unit for a Web site.

**Profiling:** Capturing individual user’s interests and needs.

**Self-Adaptive:** Ability of a Web portal to automatically adjust its presentations to perceived user’s preferences.

**Session Object:** Information items that capture characteristics and activities of a user during a Web session.

Chapter 2.27
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites: A User-Centered Approach to Web Site Development

Napawan Sawasdichai
King Mongkut’s Institute of Technology, Ladkrabang, Thailand

ABSTRACT
This chapter introduces a qualitative study of user’s information-seeking tasks on Web-based media, by investigating user’s cognitive behaviors when they are searching for particular information on various kinds of Web sites. The experiment, which is a major part of the recently completed doctoral research at the Institute of Design-IIT, particularly studies cognitive factors including user goals and modes of searching in order to investigate if these factors significantly affect users' information-seeking behaviors. The main objective is to identify the corresponding impact of these factors on their needs and behaviors in relation to Web site design. By taking a user-based qualitative approach, the author hopes that this study will open the door to a careful consideration of actual user needs and behaviors in relation to information-seeking tasks on Web-based media. The results may compliment the uses of existing quantitative studies by supplying a deeper user understanding and a new qualitative approach to analyze and improve the design of information on Web sites.

INTRODUCTION
When visiting a Web site, each user has a specific goal that relates to a pattern of needs, expectations, and search behaviors. They also approach with different modes of searching based on varied knowledge, experience, and search sophistication. This leads to differences in information-seeking strategies and searching behaviors. Since information on Web sites is traditionally structured and presented based on Web sites’ goals and contents,
it may or may not match with user goals or search behaviors.

Because of these problems, information structuring is the essence of Web design since these problems cannot be solved by the development of technically sophisticated systems alone. User search behaviors need to be studied and deeply understood in order to design systems that allow them to perform their information-seeking tasks easily, without struggle and frustration. The contents need to be authored, organized, structured, and presented to fit their needs, expectations, and search behaviors, while being able to carry out the goal of the Web site simultaneously. Both the provider and user must benefit at the same time to ensure the Web site success. As a result, user-centered design process is important in Web development to help people succeed within an information context that seeks to achieve business goals (Brinck, Gergle, & Wood, 2002).

In attempts to move toward user-centered design, many studies have been developed to establish design principles that better serve Web-based media. Among these attempts, Web usability, grounded in human-computer interaction (HCI), has currently assumed a significant role underpinning the design of many Web sites in order to maximize efficient use. Web usability studies and practices are primarily concerned with people performing a task accurately, completely, and easily. These may involve making information accessible, retrievable, legible, and readable, ensuring that all Web pages are reachable and practically navigated, or dealing with technical aspects of media interface and Web system by ensuring that all system functionality can be operated correctly and easily.

**User Research in Web Development**

User research in relation to Web site development is mostly conducted by using quantitative methods or automated programs, such as data mining and Logs File Analysis (analyze usage data), GOMS analysis (predict execution and learning time), and Information Scent modeling (mimic Web site navigation) serve different purposes. These automated methods are particularly essential to usability testing (evaluation), especially in cases where numerous users are involved since they can reveal a substantial amount of information with regard to usage patterns by representing the actual usage characteristics. Some also provide in-depth statistical analysis of usage. For example, *logs file analysis* can show overall hits, conversion rates, entrance pages, search terms, peak times, demographics, and system down-time (see Figure 1 and 2). These develop an understanding of how the Web site is being used by the actual users, which helps identify potential problems of the Web site, and may assist in suggesting a change or directing the future design (Brinck et al., 2002).

However, the limitations of these automated methods are that they cannot be employed without an existing Web site; the Web site needs to be prototyped or implemented at some level before these methods can be applied since they are intended as an analytical means rather than a generative one. More importantly, these automated methods cannot capture important qualitative and subjective information such as user preferences and misconceptions (Ivory & Hearst, 2001). They tend to yield a higher level of user data — what they do or what they do not do — but they usually fail to capture and analyze user cognitive behaviors such as their satisfaction, decision-making pattern, or reasons that underpin their needs and behaviors.

Therefore, qualitative study using nonautomated methods such as user observation, focus groups, user interviews and surveys still play an important role in Web development. These nonautomated methods can be used in the design process to capture, analyze, and conceptualize Web structure before usability evaluation takes
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

Applying both quantitative and qualitative studies can significantly improve the quality and quantity of user input and feedback that may help suggest a change or direction that better supports user satisfaction. As a result, this study investigates a relatively new area of user research in Web-based media, offering a user-centered perspective with place. They can be employed in the evaluation process as complements to the automated methods as well, in order to help capture and analyze qualitative and subjective user information that is missing from the automated methods alone.

Figure 1. An example page from logs file analysis: Visits

![Visits](image1)

Figure 2. An example page from logs file analysis example: Top entry pages

![Top Entry Pages](image2)
consideration of user goals, user modes of searching, and their search behaviors by taking a user-based qualitative approach. The study expects to expand understanding within the area of user studies, and accordingly investigates how these user cognitive factors contribute to differences in user information needs and their information-seeking behaviors on Web-based media, particularly in view of user search strategies and user search methods. Understanding within this area will contribute to the further development of information architecture and interface design.

SCOPE OF THE RESEARCH

The practices of Web site development are fundamentally concerned with two equally important and interrelated parts: (1) Web functionality and (2) Web information. In the user-centered perspective, usability, accessibility, sustainability, suitability, credibility, and usefulness of both Web functionality and Web information for its intended users are important for the Web to succeed. In most current practices, the user-centered approach is usually taken into design consideration in a general sense; for example, by conducting user studies to establish who are the intended users of the Web site (user profiles), and what do they want to achieve (user goals)? Others may perform user testing with regard to usability evaluation in terms of what is working, and what is not (usability of Web functionality). These current user-centered approaches are concerned with Web functionality rather than Web information. Furthermore, they pay considerably more attention to usability aspects while the Web information content receives less attention. Therefore, this research primarily focuses on the design of Web information, particularly in view of the importance, suitability, and usefulness of information design provided on Web sites (see Figure 3).

The study is particularly concerned with user information needs and user information-seeking behaviors; it also investigates whether the design of information provided on the Web site supports these needs and behaviors. Secondly, the research is also concerned with the suitability and usefulness of Web functionality necessary for users to gain access to the information they need. The research also looks into different types of search methods or search tools provided on Web sites to investigate whether these search methods or tools are useful and suitable to user search strategies.

When searching, each user has a different search plan: For instance, they lightly explore or seriously search and they select search methods and tools in order to easily achieve their goal. This search plan is primarily based on their search strategy, which is the scheme that generally distinguishes user search patterns. Besides search strategies, each user may use a particular search method, which is the procedure for how they actually perform their search. User search method is based on the types of search tools they choose for their search, which may include menu bar, table of contents, index, or search engine. User search strategies can range from general or less-focused search to more specific or purposeful search. Furthermore, it may change from time to time based on the current context or situation that unfolds while they are searching as well as the search results that they find or retrieve. Based on user search strategies and its results, user search methods are changed accordingly.

As a result, this research is aimed to uncover the primary factors governing or influencing these user search strategies and methods.

More importantly, in order to understand and eventually determine what types, characteristics, formats, and presentation methods for information is suitable and useful for users, the research needs to investigate the relatively new areas of user-centered approaches to Web site design: user goals and user modes of searching (see Figure 4).

Information scientists have studied user modes of searching for decades, and these ideas are well categorized by Rosenfeld and Morville (1998) in
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

Figure 3. Primary focuses of the research: The design of Web information-based on user-centered approach, particularly with regard to usefulness and suitability

Their book Information Architecture for the World Wide Web. However, since the notion of user modes of searching has never been elaborated in terms of what to expect from their differences, it needs further investigation — this becomes the primary focus of this research to uncover its substantial impact on user needs, expectations, search strategies, or information-seeking behaviors in this medium. In addition, typical user profiles are also categorized in order to determine whether these profiles exert any substantial impact on user needs, expectations, or search patterns. These profiles particularly focus on user demographic and techno-graphic data,
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

Figure 4. Primary focus of the research: User goals and user modes of searching

including prior knowledge in the content they are searching, prior experience in the particular or similar Web site interface, and sophistication in general Internet use.

RESEARCH ANALYTIC FRAME

The research begins with defining and categorizing the important elements or factors of the study, including user goals, user modes of searching, as well as Web site goals and contents. Accordingly, the research analytic frame is established to help identify potential cases for the study. User goals and modes of searching were investigated within the context of specific Web site goals to reveal common user search patterns, search strategies, and search methods associated with each case and to identify the primary problems that occur in each pattern.

User Goals

Each user has a specific goal when visiting a Web site. Different goals suggest different kinds of needs, expectations, and search behaviors, which are factors in Web usage and success. Further, users may access the same Web site with different goals at different times; moreover, they often link several goals and explore them sequentially. User goals may be categorized as follows:

- **To seek specific information:** In this category, users may engage a Web site to search for specific information that helps them to stay updated, make decisions, fulfill a specific inquiry, perform tasks, learn, or conduct research.
- **To fulfill personal interests:** Some users may engage a Web site as a resource for pleasure to fulfill a personal interest (e.g., watching a video clip or listening to music on an entertainment Web site).
- **To communicate and/or perform tasks:** Others may use Web sites as a channel for communicating or as a means for performing tasks (e.g., connecting to a community Web site or paying bills on a company Web site).

Among these three categories of user goals, the information-seeking goal is prevalent and poses the greatest problem for users. Consequently this is the primary investigative focus in this research.

User Modes of Searching

Besides user goals, users also approach a Web site with varied levels of specification of their
needs and different levels of search specification and determination, this leads to differences in information-seeking behaviors including search strategies, search methods, and selection of search tools. Some users may know exactly what they are looking for and where to find it, while others are without a clue. Since these search behaviors and user expectations vary widely, it is important to recognize and distinguish among them noting their differences.

A current study (Rosenfeld & Morville, 1998) has delineated users’ different modes of searching as known-item searching, existence searching, exploratory searching, and comprehensive searching (research). Based on Rosenfeld and Morville’s model, user modes of searching are modified and extended in this research to include topical searching which falls between existence and known-item searching. User modes of searching may be categorized as follows:

• **Exploratory searching (browsing):** Users have a vague idea of their information needs. They do not know exactly what they are hoping to find, but some may know how to phrase their question. They want to explore and learn more.

• **Existence searching:** Users have an abstract idea or concept of what they are hoping to find, but do not know how to describe it clearly or whether the answer exists at all. They want to search for what matches their idea or mental image.

• **Topical searching:** Users know what they want in general. Some may want to search for an answer to their specific question. They know what they are hoping to find, but do not know where/which categories they should look for.

• **Known-item searching:** Users know exactly what they want, and usually know where/which categories they should look for. Users’ information needs are clearly defined and have a single, correct answer.

• **Comprehensive searching (research):** Users want to search for specific information, and they want everything available regarding this information. Users’ information needs are clearly defined, but might have various or many answers.

**Users Search Behaviors**

When users are searching for information and trying to accomplish their goals, they move between two cognitive states (thoughts/decisions — with regard to their goal and interest) and physical states (interactions — concern with functions, navigation, and computer performance) with regard to information provided on each Web page.

For instance, some users may want to keep searching because they need detailed information, while others may be satisfied with only a short descriptive text presented on the first page. Some may prefer textual information, but others may feel more satisfied with visual information. These search behaviors may be identified as shown in Table 1. These cognitive and physical behaviors with regard to user search, previously observed from initial user observation, will be further used to establish a coding scheme used in an analytical process of the research.

**Web Site Goals and Contents**

While this research focuses on the relatively new areas of user studies: user goals and user modes of searching, other factors such as site contents, site goals, and site functions nevertheless play a significant role in determining the design of a Web site. Each Web site contains unique contents and goals depending on the nature of the company, institution, or individual that owns that Web site. Based on the book *Web Navigation: Designing the User Experience* (Fleming, 1998), these Web sites can be distinguished and generalized by the similarities of their goals and contents into six categories: (1) commercial Web site, (2) identity
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

Table 1. Users’ cognitive and physical search behaviors

<table>
<thead>
<tr>
<th>Users’ Cognitive Behaviors (thoughts/decisions)</th>
<th>Users’ Physical Behaviors (interactions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>— Some information is found, and they want to learn more, or want to know the details of the retrieval documents.</td>
<td>— Users keep searching in the current retrieval results.</td>
</tr>
<tr>
<td>— The intended information is found. Users’ primary information needs are fulfilled, but users are interested in finding other relevant or related information.</td>
<td>— Users keep searching by changing search strategy.</td>
</tr>
<tr>
<td>— The intended information is found. Users’ primary information needs are fulfilled. Users are ready to use information they found to take further action(s).</td>
<td>— Users record the information they found.</td>
</tr>
<tr>
<td>— The intended information is not found, or not enough to take further action(s). Users’ primary needs are not fulfilled. Users need to keep searching.</td>
<td>— Users go back to the selected (bookmarked) page or results.</td>
</tr>
<tr>
<td>— Users make a negative decision (decide not to proceed) about something according to information they found.</td>
<td>— Users give up.</td>
</tr>
<tr>
<td>— Users are satisfied. All users’ needs are fulfilled. Users are able to accomplish their goals based on the information they found.</td>
<td>— Users are not satisfied. Users’ needs are not fulfilled. Users are unable to accomplish their goal(s).</td>
</tr>
<tr>
<td>— Users are not satisfied. Users’ needs are not fulfilled. Users are unable to accomplish their goal(s).</td>
<td>— Users are not satisfied. Users’ needs are not fulfilled. Users are unable to accomplish their goal(s).</td>
</tr>
</tbody>
</table>

Web site (Web site for company or institution), (3) information Web site, (4) education Web site, (5) entertainment Web site, and (6) community Web site. However, only the first four categories, in which the problems of information-seeking tasks are primarily found, will be investigated in this study. Moreover, entertainment and community Web sites are quite different from other Web sites because of their unique goals, contents, and functions.

By simultaneously considering the three important factors of information design on
Web sites: (1) Web site goals and contents, (2) user goals, and (3) user modes of searching, an analytic frame is constructed. Different aspects of each factor are systematically combined with one another to establish prominent cases or scenarios for the study; each of which presents a unique combination of the three factors: (1) Web site goals and contents, (2) user goals, and (3) user modes of searching.

Nevertheless, these cases are not mutually exclusive; they might overlap or combine since one Web site may consist of two or more combinations (see Figure 5). As shown in Figure 6, case 1 represents the scenario in which users with

Figure 5. The research analytic frame: Generating 10 different study cases for the research
exploratory searching mode visit a commercial Web site in order to find information to make a decision. Case 2 represents a similar scenario to case 1; however, users in case 2 arrive with an existence mode of searching. Cases 3, 4, and 5 represent similar scenarios in which users visit identity (company) Web sites to find information to fulfill a specific inquiry.

However, each case has a distinctive search mode. Users in case 3 approach with an existence mode; case 4 with a topical mode; while case 5 approaches with a known-item mode. Cases 6, 7, 8, and 9 represent the scenarios in which users visit information Web sites to find information to stay updated. Though, they are assigned different modes of searching, which include exploratory, existence, topical, and known-item modes respectively. Case 10 represents a scenario in which users with a comprehensive mode of searching approach an educational Web site to find information for learning or researching a specific content. Each of these 10 cases will be investigated and analyzed to uncover similarities and differences in patterns of user information-seeking behaviors, as well as to identify user information needs, user search strategies, and user search methods associated with different user goals, modes of searching, and Web site characteristics.

**RESEARCH QUESTIONS**

The study is specifically conducted within these selected 10 cases generated from the research analytic frame shown in Figure 5 in order to find the answers to these research questions:

- What are the common patterns of user information-seeking behavior presented in each study case?
- What are the user search strategies, search methods, or selected search tools commonly found or employed in each study case?
What kinds of information do users need in each study case in terms of the types, characteristics, formats, presentation methods, quantity and quality of information?

What are the key factors in each study case that help or obstruct users to accomplish the information-seeking task?

The research findings that answer these questions will be analyzed to identify the relationships existing among user goals and user modes of searching with their information needs, search strategies, and search methods. These results will help establish the classification of cognitive factors, as well as provide an analysis framework for information design on Web sites.

RESEARCH METHODOLOGY

Research Methods

A qualitative research method is used in this study to explore the similarities and differences of user search patterns. Users’ information-seeking behaviors are observed through controlled observation, through video observation combined with protocol analysis. User profiles are also collected through a series of questionnaires. Ten scenarios are designed to create the 10 study cases originating from the research analytic frame to help the participants enter the situation and the tasks they needed to accomplish. Each scenario is embedded with a particular mode of searching, and a different search goal resulting in the performance of a task, ranging from open-ended to very specific purpose and search.

Scenario 1 explores a commercial Web site: Expedia.com. User goal is to make a decision; user search mode is exploratory searching.

Scenario 2 explores a commercial Web site: Toyrus.com. User goal is to make a decision; user search mode is existence searching.

Scenario 3 explores an identity Web site: Paris-ile-de-France.com. User goal is to fulfill a specific inquiry; user search mode is existence searching.

Scenario 4 explores an identity Web site: Apple.com. User goal is to fulfill a specific inquiry; user search mode is topical searching.

Scenario 5 explores an identity Web site: FoodTV.com. User goal is to fulfill a specific inquiry; user mode is known-item searching.

Scenario 6 explores an information Web site: TVGuide.com. User goal is to stay updated on some interesting topics; user search mode is exploratory searching.

Scenario 7 explores an information Web site: ABCNews.com. User goal is to stay updated on a specific topic; user search mode is existence searching.

Scenario 8 explores an information Web site: DiscoveryHealth.com. User goal is to stay updated on a specific topic; user search mode is topical searching.

Scenario 9 explores an information Web site: CNN.com. User goal is to stay updated on a specific topic; user mode is known-item searching.

Scenario 10 explores an information/education Web site: WebMD.com. User goal is to research and learn about a specific topic; user search mode is comprehensive searching.

Fifty participants from different cultures, all of whom were literate in English, are observed regarding how they search for information and try to accomplish the tasks defined by the scenario.
they received. The participants approached the selected Web site with unspecified and various modes of searching and searched for information with goals appropriate to the selected scenario. Ten participants are randomly selected to participate in each scenario, with each participant doing two cases or two different scenarios.

As a result, the research collects in total 100 observation cases, which consist of 10 cases for each of 10 scenarios. The participants’ interactions (physical behaviors) on Web sites are simply captured through a video recorder. Furthermore, by using protocol analysis, the participants express verbally what they think while performing tasks in order to reveal their thoughts (cognitive behaviors) and comments, which are extremely important for the analytical process.

**Analysis Methods**

Since the research data collected from participants is qualitative in nature, several methods of qualitative analysis are used in this research to carefully analyze various aspects of the data, in order to obtain integrated research findings that answer the related but different research questions on which this research focuses. Each analysis method used in the study delivers distinctive analytical results answering a specific research question. The analytical results obtained from these different analysis methods are also cross-examined in order to accumulate further findings. This collective analysis process helps to uncover the pattern of relationship that exists among various user cognitive factors, as well as to identify their substantial impact on user search behaviors and information needs.

*Thematic analysis* (Boyatzis, 1998), the process used for encoding qualitative information, is performed in this study to analyze the overall user search behaviors including user task list and process. In order to uncover the differences and similarities in user search behaviors, a thematic analysis framework with a coding scheme is designed based on an initial observation on user search behaviors. Participants’ search behaviors are captured through video and sound recording, then analyzed and encoded by using the coding scheme (see Table 2).

User search behaviors are analyzed at each Web page the user visited as cognitive behaviors (thoughts/decisions) and physical behaviors (interactions). Each behavior is encoded using the preset coding scheme. The result is the sequence of user tasks performed by each user when searching for specific information on the particular Web site as described in the scenario they received (see Table 3).

The results, displayed as the encoded information of user search behaviors, are then further analyzed and generalized to determine the common patterns of information-seeking tasks that are associated with each study case (scenario) by using a *time-ordered matrix* (Robson, 1993). The time-ordered matrix is used to systematically display the encoded information of user search behaviors in *time-ordered sequences* by presenting various types of user search behaviors, both physical and cognitive behaviors, observed in each Web page from the start to completion of the task (see Figure 7).

Color coding is also added to help identify and group the same or similar tasks together. This enables one to see the overall task list and its sequence visually and practically in order to compare the differences and similarities that occur within and across different scenarios.

In addition, the results gained from thematic analysis are eventually summarized as *procedural analysis*, which presents the common process or pattern of user search behaviors in each study case (scenario) including search methods and task descriptions (see Figure 8).

The encoded information of user search behaviors is also further transformed into Chernoff Faces (Chernoff, 1973; Wainer & Thissen, 1981) in order to further identify and compare the common patterns of search behaviors that are associ-
Table 2. The coding scheme used in thematic analysis

<table>
<thead>
<tr>
<th>Thematic Analysis: Coding Scheme</th>
<th>Physical Behaviors (Interactions)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User’s Search Behaviors:</strong></td>
<td></td>
</tr>
<tr>
<td>Cognitive Behaviors (Thoughts, Decisions)</td>
<td>Physical Behaviors (Interactions)</td>
</tr>
<tr>
<td>Some information is found and they want to learn more, or want to know about the details of the retrieval documents.</td>
<td>Users keep searching in the current retrieval results.</td>
</tr>
<tr>
<td>The intended information is found. Users’ primary information needs are fulfilled, but users are interested in finding other relevant or related information.</td>
<td>Users record the information they found.</td>
</tr>
<tr>
<td>The intended information is found. Users’ primary information needs are fulfilled. Users are ready to use information they found to take further action(s).</td>
<td>Users keep searching by changing search strategy or search methods.</td>
</tr>
<tr>
<td>Users make a positive decision (to proceed) about something according to information they found.</td>
<td>Users go back to the selected (bookmarked) page or result.</td>
</tr>
<tr>
<td>Users make a negative decision (not to proceed) about something according to information they found.</td>
<td></td>
</tr>
<tr>
<td>Users are satisfied. All users’ needs are fulfilled. Users are able to accomplish their goals based on the information they found.</td>
<td></td>
</tr>
<tr>
<td>Users are somewhat satisfied, but not completely satisfied. Users’ primary needs are fulfilled, and users are able to accomplish their goals based on the information they found. However, users still need more information to fulfill all their needs completely.</td>
<td></td>
</tr>
<tr>
<td>Users are not satisfied. Users’ needs are not fulfilled. Users are unable to accomplish their goal(s).</td>
<td></td>
</tr>
<tr>
<td><strong>Integrated Behaviors</strong></td>
<td></td>
</tr>
<tr>
<td>Initial searching</td>
<td>SR SR</td>
</tr>
<tr>
<td>Information-collecting</td>
<td>C C C</td>
</tr>
<tr>
<td>Struggling</td>
<td>A A</td>
</tr>
<tr>
<td>Decision-making</td>
<td>B B A</td>
</tr>
<tr>
<td>Satisfactory / Unsatisfactory</td>
<td>S S X</td>
</tr>
</tbody>
</table>
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

Table 3. An example analysis of thematic analysis by using the pre-designed coding scheme

<table>
<thead>
<tr>
<th>Information display on Web page</th>
<th>User’s key actions</th>
<th>User’s key speech/thoughts</th>
<th>User’s cognitive and physical behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>page 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Homepage</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Menu bar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Table of contents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Search field</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Recommend features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Well ... I want to look around first.”</td>
<td>Some information is found, and they want to learn more, or want to know about the details of the retrieval documents.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“There’re lots of categories to choose from here, but I think I should start searching by ‘Ages.’”</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Users keep searching in the current retrieval results.</td>
</tr>
<tr>
<td><strong>page 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Result Page</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Table of contents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Recommend products</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(small image + short description)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Let’s see if anything interesting is here.”</td>
<td>Some information is found, and they want to learn more, or want to know about the details of the retrieval documents.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Visual stimulation ... um ... it sounds interesting.”</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Well ... let’s see what’s in it.”</td>
<td>Users keep searching in the current retrieval results.</td>
</tr>
<tr>
<td><strong>page 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Result Page</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Small images + short descriptions</td>
<td></td>
<td>“Well ... nothing interesting here.”</td>
<td>The intended information is not found, or not enough to take further action(s). Users’ primary needs are not fulfilled. Users need to keep searching.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Maybe I should try another category to see if it has more interesting items.”</td>
<td>Users keep searching by changing search strategy.</td>
</tr>
</tbody>
</table>

different scenarios. For example, in this coding scheme, the face is used to represent the user’s information-collecting state. The bigger the face, the more information has been collected by the user. The eyes represent how much the user searches, while the eyebrows represent a struggling state (see Tables 4.1, 4.2, and Figure 9).

As shown in Figure 9, Chernoff Face Analysis reveals the patterns of user’s prominent tasks performed in each different scenario. For example, users in scenario 2, as shown in the top row, need to perform an extensive decision-making task indicated by the gray and black hair they are all wearing; in contrast to users in scenarios 4 and 9.
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

who all appear with no hair signifying that they do not perform any decision-making task at all. The analysis also visually addresses user search struggle or satisfaction clearly. As seen clearly in Figure 9, all users in scenario 9 appear with complete satisfaction while most users in scenario 4 are unsatisfied with their search.

In order to identify the patterns of user search strategies and methods commonly used in each scenario, a checklist with a sequence record (Robson, 1993) is designed to record participants’ frequency and sequence of use of various search tools available on each Web site. The recorded data is then further analyzed to identify the com-

Figure 7. An example use of time-ordered matrix used for further analyzing and generalizing the encoded information gained from the earlier thematic analysis by presenting user search behaviors in the time-ordered sequences.

*Participant N/n means scenario ‘N’/participant ‘n’.
mon patterns of user search strategies and search methods primarily used in each scenario, as well as to compare the differences and similarities of user search patterns within and across different scenarios (see Table 5).

Similarly, in order to identify the patterns of user information needs commonly found in each study case, another Checklist Record (Robson, 1993) is designed to record the frequency of use of different kinds of information available on each Web site. Different types, characteristics, formats, and presentation methods of information display that are viewed by users while performing information-seeking tasks are captured by using the Checklist Record. This process is used to analyze and identify the main types, characteristics, formats, and presentation methods of information needed by users to accomplish the given task within and across different scenarios (see Table 6).

The analysis of user profiles is collected and built upon user data acquired through a series of questionnaires provided to the research participants when the observation took place (see Figure 10). The questionnaire was designed to acquire user demographic and techno-graphic data focusing on different aspects of user experience, including users’ prior knowledge and experience in the specific content they are searching, users’ prior experience in the particular or similar Web site interface, and users’ general experience in Internet use for information-seeking purposes.

Furthermore, qualitative comparison is conducted by constructing truth tables (Ragin, 1987, 1994) to help further identify additional insights and various factors that may reveal additional
**Table 4.1. The coding scheme used in Chernoff Face Analysis**

<table>
<thead>
<tr>
<th>Chernoff Faces Analysis: Coding Scheme</th>
<th>Chernoff Faces coding system</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Users’ cognitive states and physical states</strong></td>
<td><strong>EYES</strong></td>
</tr>
<tr>
<td><strong>Initial searching states</strong></td>
<td>less often</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>more often</td>
</tr>
<tr>
<td><strong>SR</strong></td>
<td>Users keep searching in the current retrieval results.</td>
</tr>
</tbody>
</table>

| **Information-collecting states** | **HEAD** |
| | less often | 5 |
| | | 10 |
| | | 15 |
| | more often | 20 |
| **C** | The intended information is found. Users’ primary information needs are fulfill. Users are interested in finding other relevant or related information. |
| | The intended information is found. Users’ primary information needs are fulfill. Users are ready use information they found to take further action(s). |
| | Users record the information they found. |

| **Struggling states** | **EYEBROWS** |
| | less often | 0 |
| | | 5 |
| | | 10 |
| | | 15 |
| | more often | 20 |
| ? | The intended information is found. Users’ primary information needs are fulfill. Users are ready use information they found to take further action(s). |
| ! | Users keep searching by changing search strategy. |

Information about user search struggle or success, and help confirm the results from other analytical methods (see Table 7). This particular analysis is important for the research since it looks across all 100 cases simultaneously, in contrast to other analyses that examine 10 observation cases of each scenario collectively.
### Table 4.2. The coding scheme used in Chernoff Face Analysis

<table>
<thead>
<tr>
<th>Chernoff Faces Analysis: Coding Scheme</th>
<th>Chernoff Faces coding system</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Users’ cognitive states and physical states</strong></td>
<td><strong>Decision-making states</strong></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Cheek" /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Nose" /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Hair" /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Mouth" /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Mouth" /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Mouth" /></td>
</tr>
<tr>
<td><strong>Users’ cognitive states and physical states</strong></td>
<td><strong>Satisfactory states</strong></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Star" /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Star" /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Star" /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Star" /></td>
</tr>
</tbody>
</table>

- **Positive decision**: Users make a positive decision (to proceed) about something according to information they found.

- **Negative decision**: Users make a negative decision (not to proceed) about something according to information they found.

- **Positive decision**: Users go back to the previously selected or recorded (bookmarked) pages or results, and/or compare the selected pages or results side by side in case there are more than one page or result selected.

- **Negative decision**: Users are satisfied. All users’ needs are fulfilled. Users are able to accomplish their goals based on the information they found.

- **Negative decision**: Users are somewhat satisfied, but not completely satisfied. Users’ primary needs are fulfilled, and users are able to accomplish their goals based on the information they found. However, users still need more information to fulfill all their needs completely.

- **Negative decision**: Users are not satisfied. Users’ needs are not fulfilled. Users are unable to accomplish their goal(s).
Validation of Coding System

Reliability and validity of coding schemes specifically designed to use for analysis in this research is fundamentally important and needs to be examined before the study proceeds further. In qualitative research, the observer’s consistency and bias in interpreting user behaviors and using coding schemes to code events are a primary concern. As a result, to measure reliability and validity of the coding scheme and analysis methods, a second observer is invited to independently...
**Table 5. An example analysis of user search strategies and user search methods by using the Checklist and Sequence Record**

<table>
<thead>
<tr>
<th>Central tendency: mean</th>
<th>Exploring/Browsing</th>
<th>Purposeful searching</th>
<th>Auxiliary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Menus Table of content Feature items/topics list of items/topics Advertising Related items/topics Table or diagram w/link text Search field Simple search function Advanced search function Index Shortcut Site map &quot;Back&quot; button &quot;Next&quot; button &quot;See more&quot; button</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant 2/1* 2**</td>
<td>5</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Participant 2/2</td>
<td>2</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Participant 2/3</td>
<td>3 2</td>
<td>4 2</td>
<td>5</td>
</tr>
<tr>
<td>Participant 2/4</td>
<td>4</td>
<td>2 1 1</td>
<td>4 6 5</td>
</tr>
<tr>
<td>Participant 2/5</td>
<td>2</td>
<td>2 3</td>
<td>8 10</td>
</tr>
<tr>
<td>Participant 2/6</td>
<td>6</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>Participant 2/7</td>
<td>13</td>
<td>1</td>
<td>15 1 7</td>
</tr>
<tr>
<td>Participant 2/8</td>
<td>3</td>
<td>1</td>
<td>6 7</td>
</tr>
<tr>
<td>Participant 2/9</td>
<td>4</td>
<td>5 6</td>
<td></td>
</tr>
<tr>
<td>Participant 2/10</td>
<td>8</td>
<td>1</td>
<td>10 2 7</td>
</tr>
</tbody>
</table>

**Checklist and Sequence Record:** Showing the frequency and sequence of use of different search tools — Scenario 2: commercial Web site + making decision goal + existence searching mode
Table 6. An example analysis of user information needs by using the Checklist Record

<table>
<thead>
<tr>
<th>Characteristics of Information</th>
<th>Formats of Information display</th>
<th>Presentation Methods of Information display</th>
<th>Types of Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quick references, Bullet points</td>
<td>FAQs, questions and answers</td>
<td>Comparison information</td>
<td>Feature articles</td>
</tr>
<tr>
<td>Glossary, explanations</td>
<td>Stories, sequence information</td>
<td>Topical, keywords, headlines</td>
<td>Abstracts, summaries</td>
</tr>
<tr>
<td>Complete description</td>
<td>Short or brief text information</td>
<td>Full or long text information</td>
<td>Biographical references</td>
</tr>
<tr>
<td>Textual descriptions</td>
<td>Diagrams, maps</td>
<td>Images, icons, illustrations</td>
<td>Opinions, reviews and recommendations</td>
</tr>
<tr>
<td>Biographical references</td>
<td>News, reports</td>
<td>Facts, scientific information</td>
<td></td>
</tr>
</tbody>
</table>

| Participant 2/1 | 8 | 2 | 7 | 15 | 9 | 23 | 7 | 24 | 21 | YES | YES |
| Participant 2/2 | 25 | 2 | 3 | 14 | 8 | 22 | 3 | 22 | 23 | YES | YES |
| Participant 2/3 | 22 | 4 | 7 | 11 | 11 | 22 | 7 | 22 | 24 | YES | YES |
| Participant 2/4 | 22 | 4 | 3 | 15 | 7 | 23 | 3 | 23 | 24 | YES | YES |
| Participant 2/5 | 24 | 1 | 4 | 11 | 15 | 26 | 4 | 26 | 27 | YES | YES |
| Participant 2/6 | 25 | 6 | 2 | 19 | 11 | 30 | 2 | 30 | 30 | YES | YES |
| Participant 2/7 | 29 | 10 | 30 | 6 | 36 | | 35 | 36 | YES | YES |
| Participant 2/8 | 36 | 4 | 2 | 9 | 7 | 16 | 2 | 16 | 18 | YES | YES |
| Participant 2/9 | 16 | 4 | 8 | 4 | 12 | | 12 | 15 | YES | |
| Participant 2/10 | 27 | 8 | 21 | 6 | 27 | | 26 | 30 | YES | YES |

Central tendency (mean) 23.8 4.5 2.8 15.4 8.4 23.7 2.8 23.6 24.8

Remarks: Users thoughts, comments on their information needs

Many users would like to see more and bigger pictures, or some interactive displays showing the usage of product.

Some users expect to see the same or similar information to what they would see on the package of the product when they buy in a store.

Most users want to see comparison information or want a comparison tool.
interpret and code the same video data. The scripts encoded by both observers are then compared to identify the degree to which both observers agree in their interpretation and coding. This validation process is called double coding, which is perhaps the most used technique to attain sufficient reliability to proceed with analysis and interpretation (Boyatzis, 1998; Miles & Huberman, 1984).

After the double coding process is completed, the confusion matrix is constructed to show where the two observers are different in their judgment when coding the events. Agreement takes place when both observers use the same code for the same event. On the contrary, disagreement occurs when observers use different codes to code the same event. To read the confusion matrix, the scores on the diagonal from top left to bottom right indicate agreement between the two observers, while the scores off this diagonal indicate their disagreement (Robson, 1993) (see Figure 11).

As shown in Figure 11, the scores on the diagonal from top left to bottom right appear-
Table 7. The construction of truth table 1

<table>
<thead>
<tr>
<th>Causal conditions</th>
<th>Total instances among 100 cases</th>
<th>Output code: presence/absence of instance (P)</th>
<th>Output code: achieving goal–search success (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Have prior knowledge and/or experience in content?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have visited the Web site before (return user)?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Utilize different kinds of search tools?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read text or detailed information thoroughly?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>A</td>
<td>1*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Number ‘1’ indicates the presence of a causal condition or an output, and ‘0’ indicates its absence. **Code ‘n/a’ indicates that the output code for the particular row is not applicable or it cannot be identified because the instance of the causal combination on that row is absent.
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

Figure 11. An example construction of the confusion matrix showing the scores of agreement and disagreement between two observers in their judgment when coding the events

<table>
<thead>
<tr>
<th>Confusion Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer 1</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>H</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>J</td>
</tr>
<tr>
<td>K</td>
</tr>
<tr>
<td>L</td>
</tr>
<tr>
<td>M</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Note: Coding ‘A’, ‘B’, ‘C’, ‘D’, and ‘L’ indicate agreement between the two observers. However, for the coding ‘E’, there is also a score appearing off this diagonal which indicates an event of their disagreement. Note that in this particular case, both observers do not assign coding ‘F’, ‘G’, ‘H’, ‘I’, ‘J’, ‘K’, and ‘M’ to any events. Therefore, there is no score for unused coding in the matrix; this will be different from one case to another.

Then, based on the scores on the confusion matrix, the proportion of agreement, the proportion expected by chance, and the Cohen’s Kappa are respectively calculated to measure inter-observer agreement (see Figure 12). “The inter-observer agreement is the extent to which two or more observers obtain the same results when measuring the same behaviors (e.g. when independently coding the same tape).” (Robson, 1993, p. 221).

In order to assess the significance of Kappa scores, Fliess (1981) has suggested the following rules of thumb: the Kappa scores of 0.40 to 0.60 is considered “fair”; the Kappa scores of 0.60 to 0.75 is considered “good”; and the Kappa scores above 0.75 is considered “excellent.” The results obtained from this validation process show that validity of the coding schemes used in this study, in the view of inter-observer agreement, is positively strong. The Kappa scores of seven observation cases acquire “excellent” points (0.75-0.95), and the other three cases also show “good” scores (0.64-0.71).

However, the extent of agreement between two observers who use the coding scheme to code
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

The same events independently is also affected by some other factors. One primary factor may be the observer’s learning curve with regard to the coding scheme; one observer is more familiar with the coding scheme while the other observer is new and still learning to use the codes and/or interpret the events. Another important factor may be the observer’s lack of experience or direct contact with the actual participants in the prior observation when the events were recorded. This occurs when one observer attended the observation in person when the events were recorded while the other observer was absent. The observer who had experience or direct contact with the actual participants when the events were recorded will be able to capture the participants’ emotions or thoughts that are hard to detect through watching video data alone. As a result, the two observers may interpret the same user behavior differently since the first observer also makes judgments

Figure 12. An example calculation of the proportion of agreement, the proportion expected by chance, and the Cohen’s Kappa Score to measure the extent of inter-observer agreement

**Proportion of Agreement**

\[
P_0 = \frac{\text{(number of agreements)}}{\text{(number of agreements)} + \text{(number of disagreements)}}
\]

\[
P_0 = \frac{5 + 8 + 4 + 4 + 2 + 1}{25} = 0.96
\]

**Proportion of Expected by Chance**

\[
P_C = \frac{P_{1A} \times P_{2A}}{P_{1A} \times P_{2A}} + \frac{P_{1B} \times P_{2B}}{P_{1B} \times P_{2B}} + \frac{P_{1C} \times P_{2C}}{P_{1C} \times P_{2C}} + \frac{P_{1D} \times P_{2D}}{P_{1D} \times P_{2D}} + \frac{P_{1E} \times P_{2E}}{P_{1E} \times P_{2E}}
\]

\[
P_C = \frac{0.26 \times 0.33}{0.26 \times 0.33} + \frac{0.30 \times 0.30}{0.30 \times 0.30} + \frac{0.15 \times 0.15}{0.15 \times 0.15} + \frac{0.11 \times 0.07}{0.11 \times 0.07} + \frac{0.03 \times 0.03}{0.03 \times 0.03} = 0.1969
\]

**Cohen’s Kappa**

\[
K = \frac{P_0 - P_C}{1 - P_C}
\]

\[
K = \frac{0.96 - 0.1969}{1 - 0.1969} = 0.7631 = 0.95
\]

\[
K = \frac{0.96 - 0.1969}{1 - 0.1969} = 0.8031
\]
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

based on experience with the actual participant. These are factors that may play an important role in the inter-observer agreement.

ANALYSIS OF RESEARCH DATA

Patterns of User Search Behaviors

The research provides a new perspective on design considerations for a Web site by incorporating requirements from both Web site (client) intentions and user goals. The results from this study in which user goals and their modes of searching were investigated simultaneously with Web site goals to reveal common search patterns associated with each case and significantly show that the patterns of user search behaviors are uniquely different depending on their goals and current modes of searching. Even though each user performed his/her task in isolation and in his/her own way, similar search patterns appeared based on a shared goal and/or the same mode of searching. Different search patterns were associated with different user goals and modes of searching, as well as Web site intentions.

In this research, user search behaviors are primarily analyzed by using the thematic analysis (Boyatzis, 1998) with time-ordered matrix (Robson, 1993) (see Table 3 and Figure 7), along with procedural analysis (see Figure 8), and Chernoff Faces (Chernoff, 1973, Wainer and Thissen, 1981) (see Figure 9), to uncover the patterns of user tasks in each scenario, while Checklist and Sequence Record (Robson, 1993) (see Table 5) is used to identify the types of user search strategies and methods. The analyses (see an example in Figure 7) show that users who begin with the same goal will perform their search similarly in terms of what tasks are necessary to reach the goal. However, if they use different modes of searching, which depend mainly on how precisely they know what they want, they will have different search strategies and consequently choose different kinds of

Figure 13. User goals and user modes of searching, the main factors regulating user search behaviors and the resulting search patterns
search methods even though they begin their search with the same goal. Therefore, based on these research findings, user goals and modes of searching are the main mechanisms that play an important role in determining user behaviors and the resulting search patterns.

While user goal is the main factor regulating their task description, user mode of searching provides the major impact on search strategies and search methods. User goals determine the different tasks they need to perform to achieve their goal. Simultaneously, user modes of searching influence their search strategies, or the plans of their search, determining how hard or easy the search can be and how much time is spent on their search, which accordingly results in selecting different search methods based on their search strategies (see Figure 13).

The analyses are collectively done on 10 observation cases of each of 10 scenarios which are systematically fabricated according to the research analytic frame previously demonstrated, in order to uncover the patterns of similarities or differences of user search behaviors. Based on research findings, the participants in scenarios 1 and 2 share the same search goal. As a result, even though they have different modes of searching and perform their tasks on different Web sites, the patterns of their task descriptions are very alike. Likewise the participants in scenarios 3, 4, and 5, or the participants in scenarios 6, 7, 8, and 9 have different modes of searching on different Web sites but have very similar tasks. Each group of these participants who share the same search goal, perform their tasks similarly although they do not share the same search modes and they visit different Web sites.

**Patterns of User Search Strategies and Methods**

When performing their search, individuals need a search plan — how to perform their search and which kinds of search tools to use in order to easily achieve their original goal. This plan is different from person-to-person based on user search strategy. However as stated earlier, one’s search strategy is influenced by one’s mode of searching (see Figure 13).

As a result, when visiting a Web site, individuals who arrive with different modes of searching will form different intentions as well as plan different search strategies, and accordingly perform their search in different ways to reach the same goal. Search strategies range from a general or less objective search (browse) to a more specific or purposeful search. These are directly proportional to user modes of searching which range from open-ended to specific search. They may also plan a fast or slow search based on time available and the urgency of their need. Users who plan a slow search usually want to record their search results by using the bookmark function or simply print out the results of retrieval pages for later use. Search strategy may change from time to time in accordance with modes of searching, which are based on the current context or situation that unfolds while they are searching as well as in response to the search results they find or retrieve.

While search strategy is the scheme that generally characterizes user search patterns, search method is the procedure for how they actually perform their search. This concerns the types of search tools chosen for use in their search. These search tools may include a menu bar, table of contents, index, site map, shortcut, search engine, and so forth. Users select from a combination of these search tools to perform their search based on their strategy. For example, users who are in the mode of exploratory searching (open-ended search), will likely plan a slow and general search (search strategy) and explore by browsing the menu bar and table of contents (search method). On the other hand, users who are in the mode of known-item searching (specific search), will
usually plan a fast and purposeful search (search strategy) and comfortably use the index or shortcut (search method) to pursue their search.

The analyses (see an example in Table 5) show that users who begin with the same mode of searching have similar search strategies and choose similar methods. Based on research findings, the participants in scenarios 1 and 6 begin their search with the same mode of searching. As a result, even though they have different search goals and perform their tasks on different Web sites, they choose similar search strategies and methods. Each group of participants who share the same mode of searching chooses similar search strategies and methods although they do not share the same search goals and they visit different Web sites. However, note that even though the participants in scenarios 2, 3, and 7 have different modes of searching from the participants in scenarios 4 and 8, they also share similar search strategies and methods.

Patterns of User Information Needs

In this research, user information needs are primarily analyzed by using the Checklist and Sequence Record (Robson, 1993) (see Table 6) to identify the types, characteristics, formats, as well as quality and quantity of information preferred or needed by users to fulfill their original goals. Besides having different search strategies and methods, users also have different needs for information that can fulfill their goals. The findings demonstrate that user goals, modes of searching, and prior knowledge and experience in the contents they search are collectively the main mechanisms that play an important role in determining their information needs. Consequently, each user who comes to visit a Web site with a different goal, mode of searching, and prior knowledge and experience will need different kinds of information in order to fulfill his/her goal. Information provided on a Web site may be categorized based on various aspects of information including the characteristics of the information; formats and presentation methods of information display; types of information; as well as quality and quantity of given information.

Information characteristics differ widely including quick reference information such as short or brief information organized and presented using bullet points; frequently asked question (FAQ); glossary or explanation; procedural guideline or step-by-step instruction; comparison information; recommendation; sequential story or report; and complete description. Based on the research findings, user need for different characteristics of information is influenced by their different goals.

Types of information can be categorized into three different groups. The first group includes information that comes from personal or expert opinion, critique, review, or recommendation such as an editor’s choice or customer’s review. The second group may include the information that is collected from news or reports such as today’s news or weekly reports, and the last group includes the information that presents the facts or scientific information for any given topic or item. Similar to user need for different information characteristics, the research findings demonstrate that user information needs for different types of information are influenced by different user goals.

Formats of information display range from an abstract level, including keyword; topic or headline; abstract; summary, to the detailed level including brief/short text or information, full/long text or information, reference, and bibliography. Based on the research findings, user needs for different formats of information display are influenced by different modes of searching. Presentation methods for information display range from textual to visual presentation, including textual or descriptive information (text), diagram, matrix,
A Qualitative Study in User’s Information-Seeking Behaviors on Web Sites

table, icon, image, illustration, or combinations of these methods. The research findings show that information needs for different presentation methods of information display are influenced by the type and character of the site contents.

Quality and quantity of information range from shallow information, which is usually not enough to help individuals to take further actions, to a variety of deep and detailed information. Based on the research findings, user information needs for different levels of quality and quantity of given information are influenced by various factors related to both the user and the Web site simultaneously. These factors include user goal; modes of searching; prior knowledge and experience in the contents they search; as well as the characteristics of Web contents — simple, everyday topics, or complicated, hard topics.

Discussion on Relevant Factors for User Search Success

Even though the accuracy of a search engine is one of the most recognized factors determining user search success or failure, qualitative comparison conducted by constructing truth tables and applying Boolean algebra method (Ragin, 1987, 1994) demonstrates that there are other relevant factors that play an important role to influence user search success or struggle. These influencing factors derive from both user profiles and behaviors as well as Web site attributes.

The qualitative comparison method is used in this study to examine among cases the combinations of causal conditions that help produce the positive outcome (users achieve the goal). These causal conditions include user prior experience in the content and Web site interface, their behaviors while searching, and several Web site attributes. Two truth tables (truth table 1 [see Table 7], and truth table 2 [see Table 10]) are constructed from observation data, which is recorded into nominal-scale and represented in binary form, to display different combinations of values on the conditions and outputs.

A presence-absence dichotomy is used in the tables to specify what outcomes and causal conditions are either present or absent in each observation case. Code number 1 indicates the presence of a condition or an output; code number 0 indicates the absence (see Tables 7 and 10). Truth table 1, as demonstrated in Table 7, is constructed to examine the causal conditions of user prior experience and search behavior and identify the combinations that significantly contribute to user search success.

With uppercase letters indicating presence and lowercase letters indicating absence of a particular causal condition shown in the combination, the data on user search success (S) from truth table 1 can be represented in the Boolean equation as follows:

\[
S = ABCD + ABCd + ABcD + AbCD + AbcD + abCD + AbCd + aBcD + Abcd + abcD + abCd
\]

This equation for S (search success) shows 11 primitive combinations of causal conditions that help users to achieve their goal. In order to simplify these primitive expressions, the concept of Boolean minimization is used. The most fundamental of Boolean minimization rules is (Ragin, 1987):

*If two Boolean expressions differ in only one causal condition yet produce the same outcome, then the causal condition that distinguishes the two expressions can be considered irrelevant and can be removed to create a simpler, combined expression (p. 93).*

The Boolean minimization process is conducted in a bottom-up fashion until no further stepwise reduction of Boolean expression is possible. This process is applied to the primitive...
Table 8. Boolean minimization process applied to the primitive expressions from truth table 1

<table>
<thead>
<tr>
<th>Minimization: Step 1</th>
<th>Minimization: Step 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCD combines with ABCd to produce ABC</td>
<td>ABC combines with AbC to produce AC</td>
</tr>
<tr>
<td>ABCD combines with ABcD to produce ABD</td>
<td>ACD combines with AcD to produce AC</td>
</tr>
<tr>
<td>ABCD combines with AbCD to produce ACD</td>
<td>ACD combines with AcD to produce AD</td>
</tr>
<tr>
<td>ABCd combines with AbCd to produce ACd</td>
<td>AbC combines with Abc to produce Ab</td>
</tr>
<tr>
<td>AbcD combines with abCD to produce bCD</td>
<td>AbC combines with acD to produce bC</td>
</tr>
<tr>
<td>AbcD combines with abCD to produce AbC</td>
<td>AbC combines with acD to produce bD</td>
</tr>
<tr>
<td>AbCD combines with AbCD to produce AcD</td>
<td>AbC combines with beC to produce bD</td>
</tr>
<tr>
<td>AbCD combines with AbCD to produce AbC</td>
<td>AbC combines with beC to produce bD</td>
</tr>
<tr>
<td>AbCD combines with AbCD to produce Abc</td>
<td>AbC combines with BeC to produce bD</td>
</tr>
<tr>
<td>AbCD combines with AbCD to produce abC</td>
<td>AcD combines with acD to produce cD</td>
</tr>
<tr>
<td>AbCD combines with AbCD to produce AbC</td>
<td>AcD combines with BeC to produce cD</td>
</tr>
<tr>
<td>AbCD combines with AbCD to produce bCD</td>
<td>BeD combines with beC to produce cD</td>
</tr>
<tr>
<td>AbCD combines with AbCD to produce bCD</td>
<td>BeD combines with beC to produce cD</td>
</tr>
<tr>
<td>AbCD combines with AbCD to produce bcD</td>
<td>bCD combines with beC to produce bD</td>
</tr>
<tr>
<td>AbCD combines with AbCD to produce bcD</td>
<td>bCD combines with beC to produce bD</td>
</tr>
</tbody>
</table>

Table 9. Prime implicant chart showing coverage of original terms by prime implicants

<table>
<thead>
<tr>
<th>Primitive Expressions</th>
<th>ABCD</th>
<th>ABCd</th>
<th>ABcD</th>
<th>AbCD</th>
<th>AbcD</th>
<th>abCD</th>
<th>AbCd</th>
<th>abcD</th>
<th>abCd</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ab</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bC</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bD</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cD</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABD</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Acd</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
expressions derived from truth table 1 as demonstrated in Table 8. With the Boolean minimization process applied, the reduced expressions (prime implicants) on user search success (S) from truth table 1 can be represented in the simpler equation as follows:

\[ S = AC + AD + Ab + bC + bD + cD + ABD + Abd + Acd \]

Then, the final step of Boolean minimization is conducted by using the prime implicant chart (see Table 9) to map the links between nine prime implicants (see the second equation previously shown) and 11 primitive expressions (see the first equation). This process helps to eliminate redundant prime implicants in order to produce a logically minimal number of prime implicants which cover as many of the primitive Boolean expressions as possible.

Eventually, with the final process of Boolean minimization applied, the final equation (S) from truth table 1 demonstrates six combinations of causal conditions that produce the positive outcome (user search success) as follows:

\[ S = AC + AD + Ab + bC + bD + cD \]

This final equation significantly demonstrates the result showing that causal condition ‘A’ (users have prior knowledge and/or experience in the content), condition ‘C’ (users utilize different kinds of search tools), and condition ‘D’ (users read text or detailed information thoroughly) are the important variables that help users to achieve their goals.

Contrary to the traditional view on user experience with Web site interface (first-time versus return users), the result shows that causal condition ‘B’ (users have visited the Web site before) is not the primary factor contributing to users’ accomplishment in their search.

In addition, the second truth table (truth table 2) is constructed to examine the impact of various causal conditions including user prior knowledge in the contents they search (condition ‘A’) and prior experience in Web interface (condition ‘B’), combined with different causal conditions from various Web site attributes (see Table 10). These variables include condition ‘E’ (Web site provides different approaches to content classification), condition ‘F’ (Web site has well-organized search retrieval results), and condition ‘G’ (Web site provides search tips or examples).

With all processes of Boolean minimization applied, the final Boolean equation (S) from truth table 2 demonstrates six combinations of causal conditions that produce the positive outcome (user search success) as follows:

\[ S = AeF + AEFg + AEfG + BeFg + bEFg + Abefg \]

Significantly, this final equation derived from truth table 2 also confirms that causal condition ‘A’ (users have prior knowledge and/or experience in the content) is the important variable that helps users to achieve their goals. Besides the condition ‘A,’ the Web site variables that have significant impact on user search success include condition ‘E’ (Web sites that provide different approaches to content classification) and condition ‘F’ (Web sites that have well-organized search retrieval results). The result also shows that condition ‘B’ (users have visited the Web site before), and ‘G’ (Web sites provide search tips or examples) have less impact on user search success compared with other variables. These analytical results as well as others are further summarized and synthesized, in order to develop explanatory frameworks of user search behaviors and needs, as well as to establish classifications of substantial user factors and analytical frameworks to evaluate information design on Web sites.
Table 10. The construction of truth table 2

<table>
<thead>
<tr>
<th>Causal conditions</th>
<th>Total instances among 100 cases</th>
<th>Output code: presence/absence of instance (P)</th>
<th>Output code: achieving goal–search success (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td>Have prior knowledge and/or experience in content?</td>
<td>Have visited the Web site before (return user)?</td>
<td>Provide different approaches to content classification?</td>
<td>Have well organized search (retrieval) results?</td>
</tr>
<tr>
<td>1*</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Number ‘1’ indicates the presence of a causal condition or an output, and ‘0’ indicates its absence. **Code ‘n/a’ indicates that the output code for the particular row is not applicable or it cannot be identified because the instance of the causal combination on that row is absent.
CONCLUSION

This investigation demonstrates that a user-centered approach can improve information design on Web-based media through study of various factors, especially user cognitive factors including user goals and modes of searching, to identify the corresponding impact of these factors on information and functional needs in terms of user behaviors. As an attempt to solve the problems of information-seeking tasks in Web-based media, the research is successful in providing a new perspective on Web site design considerations by strongly taking a user-centered approach to incorporate a careful consideration of actual user needs and behaviors together with requirements from a Web site.

By conducting extensive qualitative research on user study in relation to search needs and behaviors on Web sites as well as employing various analytical methods to uncover different aspects of the research data, the study answers the research questions. The common patterns of user information-seeking behavior, user search strategies and methods, as well as user information needs presented in different cases are revealed. These valuable findings will be further synthesized to develop frameworks and classifications.

Deeper understanding of these various factors, especially user cognitive factors, may complement the use of existing analytical or design methods such as task analysis and scenario-based design, by helping Web developers to recognize the important factors that may be subtle or previously unidentified yet substantially affect user task performances. By recognizing these elements, Web developers can identify the useful and appropriate functions and/or information to include in each particular case, in order to support user needs and task performances and eventually promote their satisfaction.

REFERENCES


Chapter 2.28
Developing and Validating a Measure of Web Personalization Strategy

Haiyan Fan
Texas A&M University, USA

Liqiong Deng
University of West Georgia, USA

ABSTRACT
An important theoretical undertaking in personalization research is to identify the structure of the multidimensional construct of personalization and to operationalize them in measurable terms. The purpose of this study was to develop and validate measurement scales for personalization by identifying four distinctive personalization archetypes and hypothesizing their respective relationships with different cognitive and affective outcomes. This effort was successful in several respects. New scales for measuring personalization strategies were developed based on the definitions of personalization archetypes (architectural, instrumental, social and commercial), which were in turn derived from an extensive review of multidisciplinary studies. A lab experiment with 229 student subjects was conducted to explore the structure of the personalization construct and to validate the instrument using four Web site stimuli representing four types of personalization strategies. The new measures were found to have strong psychometric properties and exhibit significant linkages with respective cognitive and affective outcomes as theoretically postulated.

INTRODUCTION
In e-commerce and mobile commerce, personalization has been recognized as an important element in customer relationship and Web strategies. Personalization is commonly treated as an independent variable that influences Web usage outcomes such as customer experience (Novak, Hoffman, & Yung, 2000), Web site usability (Agarwal & Venkatesh, 2002; Palmer, 2002), and
Developing and Validating a Measure of Web Personalization Strategy

customer churning behavior (Chen & Hitt, 2002). However, the extant research on personalization is constrained by the shortage of high-quality measures for the personalization construct (Fan & Poole, 2006). Due to the inadequacy of existing personalization measures to capture all dimensions of personalization consumers actually value (Chen & Hitt, 2002), empirical studies have often failed to support the hypothesized association between personalization and the behavioral outcomes (Chen & Hitt, 2002; Novak et al., 2000).

In responding to the need for better measures of personalization, this study aims to develop and validate measures for four archetypes of personalization.

Personalization can be generally defined as “changing the functionality, interface, information access and content, or distinctiveness of a system to increase its personal relevance to an individual or a category of individuals” (Blom, 2000, p. 313). While there exist a set of narrow, specialized definitions of personalization that are built around varied goals and techniques of personalization in the literature, Blom’s definition of personalization provides an integrated view of personalization that centers on the general goal of increasing personal relevance without limiting the personalization concept to particular motivations, context or techniques (Fan & Poole, 2006).

In developing measures for personalization, we focus on four personalization archetypes: architectural, instrumental, social and commercial personalization, which are theorized to be four distinctive personalization strategies that elicit different cognitive and affective outcomes (Fan & Poole, 2006). While the existing personalization categories, such as Rossi et al.’s three personalization categories pertaining to link, navigation structure, and navigation context, are largely concerned with the system-level personalization mechanisms (Rossi, Schwabe, & Mattos Guimares, 2001; Wu, Im, Tremaine, Instone, & Turoff, 2003), Fan and Poole’s personalization archetypes provide a high-level framework for the design of personalization by delineating different philosophies concerning user-centered motivations behind personalization and their corresponding strategies of personalization. Developing valid measures for those four personalization archetypes will not only facilitate the investigation of cognitive and affective aspects of using the personalization system, but also promote the accumulation of personalization research results on the basis of a common measure of personalization. In addition to the significant theoretical value, developing and validating such measures also has great practical value for both personalization technology vendors and user organizations.

In the rest of this article, first we review prior literature on personalization, followed by a discussion of four personalization archetypes and proposition of hypotheses for their individual influences on cognitive and affective aspects of Web usage. Then, we present the method and procedure of developing the multi-item measurement scales of the four personalization archetypes. Next, we discuss the research design of an empirical study to test and validate the measurement model. After that, the results of data analysis are reported and interpreted. Finally, limitations of the study are addressed along with future directions.

REVIEW OF PRIOR LITERATURE

As an important social phenomenon that brings considerable economic values (Davenport & Beck, 2001; Pine & Gilmore, 1999), personalization has been studied extensively in e-commerce research. Considerable research efforts have been devoted to classifying different mechanisms for implementing personalization. Rossi et al. (2001) made a distinction between base information and behavior, what the user perceives and how the user perceives. Rossi et al.’s framework mainly deals with system-level elements such as personalization for links, navigation structure, and navigation context. Wu et al. (2003) classified personalization
Developing and Validating a Measure of Web Personalization Strategy

on e-commerce Web sites into five categories: control, content, link, customized screen design, and anthropomorphic personalization. Wu et al.’s personalization categories fit a two-by-two framework with implicit vs. explicit personalization on one dimension and Web content vs. Web interface on the other dimension. Terziyan (2001) defined three levels of personalization: service, content, and dynamic personalization. Although these findings have shed light on the practice of “how to do personalization,” there is still a lack of understanding regarding “how personalization can be done well.”

To guide the design of personalized systems, user-centered perspectives of personalization are needed to develop a vision of what personalization could be, which articulates the goals of personalization and criteria for evaluating the goal attainment by personalization. In response to this need, we developed a high level framework for classifying personalization. This framework delineates fundamental assumptions and user-centered motivations underlying personalization in the literature and relates them to design strategies for developing personalized systems. In addition to its significance for personalization design, the high-level framework of personalization is built around the motivations and goals of personalization from a user perspective, and thus provides an effective way to conceptualize and operationalize personalization dimensions that the user values. This has significant implications for studying the effects of Web personalization features on customer-related Web usage outcomes.

Our personalization framework differentiates four general personalization archetypes extracted from the personalization literature of several fields. Representative of four distinct schools of thought on the underlying motivations and goals of personalization, the personalization archetypes not only offer an analysis of possible design choices for personalization, but also provide a structure for defining general personalization strategies and examining their respective influences on user responses. The next section describes the concepts and strategies of four personalization archetypes in the Web site context, as well as the hypothesized relationships between different archetypes of personalization strategy and the cognitive or affective aspects of Web usage outcome.

FOUR ARCHETYPES OF PERSONALIZATION STRATEGY

Different schools of thought can be discerned within the diverse personalization literature. To capture the characteristic features of these logically consistent approaches to thinking about personalization, we conducted a multiparadigm review of the literature on personalization. During the review process, we utilized two metatriangulation techniques discussed by Lewis and Grimes (1999) to uncover paradigmatic disparity and complimentary. Metatriangulation is a theory-building strategy for exploring the assumptions of divergent theoretical views and gaining insights into the multiple paradigms (Lewis & Grimes, 1999). First, we used the paradigm bracketing technique to differentiate and articulate various sets of assumptions of alternative paradigms of personalization. Second, we employed the technique of paradigm bridging to identify “transition zones” (Lewis & Grimes, 1999), where paradigmatic boundaries become fuzzy and new views permeating across paradigms are synthesized.

The sample for the multiparadigm review includes 86 journal articles, 35 books or book sections, 13 conference papers, and 8 Web references, obtained from the electronic library.
Developing and Validating a Measure of Web Personalization Strategy

databases using the keywords “personalization,” variants of the same word stem, and related terms such as “customization,” “adaptation,” “individuation,” “consumer-centric,” and “one-to-one relationship.” The sample spans six general areas in which personalization has been studied: marketing/e-commerce, computer science/cognitive science, architecture/environmental psychology, information science, and social sciences, including sociology, anthropology and communication. Utilizing the paradigm bracketing technique, we distinguished disparate views of personalization by: 1) conducting a detailed analysis of each personalization study in the sample, 2) identifying and categorizing the different viewpoints on the motives or goals of personalization which match the requisite diversity in the sample, and 3) writing a paradigm account for each different viewpoint resulting from step 2.

During step 2 of paradigm bracketing, we relied on Weber’s ideal type theory in defining different views of personalization in the form of personalization archetypes. This conceptual apparatus Weber defined as the ideal type is an abstraction of essential features of a particular social or economic phenomenon in the purest form possible (Ringer, 1997; Rogers, 1969; Weber, 1968). The ideal type theory is useful for studying personalization for two reasons. First, it is important to study each of the distinct schools of thought on personalization in its “pure form” so as to capture their respective central characteristics. According to Weber (1968), an ideal type of personalization is an abstraction constructed from the accentuation of certain features of reality relevant to one theoretical view of personalization. The relevant features are put together into a consistent, logic construct and separated from other features that are only visible through other views of personalization (Weber, 1968). The resulting ideal type of personalization is thus a “pure” type in a logical sense where “the real indistinctness and ambiguity” are set aside and its relevant features are “perfect on logical grounds” with the “merit of clear understandability and lack of ambiguity” (Weber, 1968). Therefore, the ideal type theory provides a methodology for analyzing the typical or “logically consistent” features of different views of personalization. Second, there is a strong need to establish a common frame of reference against which the current practice of personalization can be evaluated. The ideal types describe “objectively possible” courses of action (Rogers, 1969) and can serve as guidelines for conducting and evaluating personalization systems in light of alternative approaches.

As a result, four archetypes of personalization were distilled from the literature: the architectural, social, instrumental and commercial personalization (Fan & Poole, 2006). Each archetype represents a different philosophy concerning the motivation behind personalization and what personalization tries to accomplish. Each archetype also implies a different strategy for personalization, different means for carrying out this strategy, and different user modeling techniques. Finally, each archetype implies different criteria for evaluating personalization systems.

Architectural personalization is defined as the construction of the digital environment to create a pleasant user space and a unique experience for the user through arrangement and design of digital artifacts in a way that meets the user’s needs and reflects his or her style and taste. Because architectural personalization is concerned with building digital environments, it relates particularly to the interface aspect of the system. The goals for personalization in this view are two-fold: (1) to create a functional and delightful Web environment that provides aesthetic value and reflects the user’s personal style; and (2) to help the user cultivate a sense of personal and social identity within the space (Becker, 1977).

Instrumental personalization refers to the utilization of information systems to enhance efficiency and personal productivity by providing, enabling and delivering useful, usable, user-friendly tools in a way that meets the user’s
Developing and Validating a Measure of Web Personalization Strategy

situated needs. Instrumental personalization focuses on the functionality of the system. Its goal is to support users in accomplishing their goals. Unlike architectural personalization, in which function and form balance each other, instrumental personalization emphasizes functionality and usability and treats aesthetics as a secondary consideration to be addressed once instrumental standards are met.

Social personalization can be defined as the personalization via the mediation of interpersonal relationships and utilization of relational resources to facilitate social interactions by providing a convenient platform for people to interact with others in a way that is compatible with the individual’s desired level of communality and privacy. The motivation behind social personalization is to personalize by fulfilling the user’s needs for socialization and a sense of belonging. The goal is two-fold: (1) to enhance the effectiveness of interpersonal interactions, and (2) to help generate “social capital” (Wellman, 2002) by providing new opportunities for strengthening social relationships and maintaining social networks. Applications amenable to relational personalization vary greatly in size and complexity. They can be as simple as providing an “e-mail to a friend” button to notify others of one’s flight schedule after booking tickets online or as complicated as a conglomeration of online information portal and activity center in a “Digital City” that engages residents or visitors (Toru, 2002).

Commercial personalization is the differentiation of product, service, and information to increase sales and to enhance customer loyalty by segmenting customers in a way that efficiently and knowledgeably addresses each user or group of users’ needs and goals in a given context. One of the most important human activities is the consumption of goods and services. The motivation of commercial personalization is to fulfill users’ material needs and thus contribute to their psychic welfare (Douglas & Isherwood, 1979). It primarily focuses on the content of the system and assumes that product, service and information of high relevance to the consumer yield a satisfying shopping experience and loyal adherence to the Web site, as well as the organization behind it. There has been much evidence that commercial personalization, by providing highly relevant products/services/information to the customers, promotes long-term customer relationship building and creates high customer retention called Web site “stickiness” in the context of e-commerce (Liu & Arnett, 2000; Zott, Amit, & Donlevy, 2000). The primary method to realize commercial personalization is segmentation. Commercial personalization is ultimately effective only to the extent that the offerings provide value to the target market segments by differentiating the product, service and information provided.

After the four personalization archetypes were identified through paradigm bracketing, we then conducted paradigm bridging, which involves the identification of paradigmatic connections, similarities, and complementarities among the four perspectives of personalization. A close examination of the four personalization ideal types reveals that the four types can be further classified into a 2x2 grid (see Table 1). First, the perspectives can be differentiated in terms of their associations with extrinsic or intrinsic motivation. The instrumental and commercial personalization archetypes support extrinsic motivation. They emphasize task performance and commercial transactions and hence are employed to enhance the usability and utility of information system for specific goal/task achievement; while architectural and social personalization archetypes attempt to support intrinsic motivation by providing self-fulfilling aesthetic and socio-emotional values for users to obtain pleasure or enjoyment from the activity itself.

The four personalization archetypes can also be differentiated in terms of the basic premise of use, whether the user primarily engages the system as an individual or through an interaction. Both architectural and instrumental personalization are
concerned with individual use of an artifact, be it a building, an information system or a Web site. Design emphasis is on individual's interaction with the artifact. On the contrary, relationships among multiple entities and the management of the relationships are of paramount importance in relational and commercial personalization.

**Effects of Personalization**

The personalization archetype scheme implies that no single standard or approach to personalization is “the best.” Each archetype employs different criteria for evaluating how well the system succeeds in delivering the desired effect. Based on our paradigm bridging of the four personalization archetypes, the effects of different personalization archetypes can be differentiated in terms of the extrinsic or intrinsic motivation they intend to support.

Commercial and instrumental personalization, predominantly used for information retrieval, transaction processing, and content management, belong to the class of productivity/utility applications. They are utilitarian-oriented, the goal of which is to achieve specific objectives or tasks. The idea behind instrumental personalization is that users will find systems that are designed and tailored to realize their particular goals or tasks as more relevant and hence useful. For commercial personalization, customizable or customized products/information/services are more useful because they meet individual needs. Because utilitarian use is goal-driven and emphasizes extrinsic motivation, useful content and functionality, as well as usability are given priority in design. Instrumental personalization that focuses on providing usable, useful and user-friendly tools is also expected to relate to the ease of use aspect of Web site design. Therefore, in conceptualizing the effects of commercial and instrumental personalization, we adopt the constructs of perceived usefulness and perceived ease of use in Davis’ TAM model (Davis, 1989), which is widely used to evaluate user perceptions of utilitarian systems under extrinsic motivation. Although the TAM model was originally developed to predict the adoption of workplace technology, it has also been successfully applied to the context of e-commerce. Studies have demonstrated the importance of TAM model in predicting intentions to adopt or revisit the Web site and online shopping behaviors (Gefen, Karahanna, & Straub, 2003; Gefen & Straub, 2000; Koufaris, 2002; Shang, Chen, & Shen, 2005; van der Heijden, 2000, 2003). Even for the Web site users who are highly computer

---

**Table 1. Personalization design paradigms**

<table>
<thead>
<tr>
<th>Architectural</th>
<th>Instrumental</th>
</tr>
</thead>
<tbody>
<tr>
<td>form and function</td>
<td>functionality and usability</td>
</tr>
<tr>
<td>Social</td>
<td>Commercial</td>
</tr>
<tr>
<td>Meaning</td>
<td>content</td>
</tr>
</tbody>
</table>

Intrinsic Motivation Extrinsic Motivation
literate, the perceived usefulness and ease-of-use of a Web site relative to the competing Web sites may still greatly influence the users’ intention to return to that Web site (van der Heijden, 2000). Therefore, we consider perceived usefulness and ease-of-use as important outcome variables of Web personalization, and formally posit the following hypotheses:

**H1:** The instrumental personalization will positively influence the perceived usefulness of the Web site.

**H2:** The commercial personalization will positively influence the perceived usefulness of the Web site.

**H3:** The instrumental personalization will positively influence the perceived ease of use of the Web site.

While instrumental and commercial personalization archetypes are generally utilitarian-oriented and focus on extrinsic motivation, architectural personalization and social personalization are closely associated with intrinsic motivation. The latter two are used primarily for creating an attractive Web environment, a comfortable, interactive social network, and a sense of psychological and social well-being. The design focus for this category centers on the enjoyment and entertainment aspect of the user experience. Hence, a balance between form and function, as well as the meaning of using the system, is emphasized. Architectural personalization affords an intrinsically pleasant experience of creating a personalized digital space through the arrangement and design of digital artifacts in a way that reflects the user’s individual style and taste and meets the user’s needs for easy navigation, intelligent presentation and aesthetic delight. To investigate the effects of architectural personalization, we employ two intrinsic motivation variables, perceived enjoyment and ease of use (Atkinson & Kydd, 1997). Davis et al. define perceived enjoyment as “the extent to which the activity of using the computer is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated” (Davis, Bagozzi, & Warshaw, 1992, p. 1113). Therefore, by definition, perceived enjoyment is largely concerned with fulfillment of intrinsic motivation, which is based on emotionally satisfying experience with the system for its own sake but bears no relation to task objectives external to the user-system interaction (Van der Heijden, 2004). Perceived ease of use has been shown to contribute to both extrinsic and intrinsic motivations (Atkinson & Kydd, 1997; Davis, 1989; Lee, Cheung, & Chen, 2005). On the one hand, perceived ease of use can improve user’s task performance, which is critical for extrinsic motivation (Davis, 1989), and on the other hand, perceived ease of use can enhance user’s self-efficacy, which is an important component of intrinsic motivation (Bandura, 1977, 1982; Lee et al., 2005). Due to the focus of architectural personalization on fulfilling user’s intrinsic motivation by providing aesthetically pleasant, easy-to-navigate Web interface, we suggest the use of architectural personalization strategy will improve user’s perceived enjoyment and ease of use of the Web site. Therefore, formally we have:

**H4:** The architectural personalization strategy will positively influence the perceived enjoyment of the Web site.

**H5:** The architectural personalization strategy will positively influence the perceived ease of use of the Web site.

The social personalization strategy personalizes one’s world by creating a unique Web of social relationships. It seeks to provide a common, convenient platform for interpersonal communication and community building that emphasizes design on the basis of what Preece terms
Developing and Validating a Measure of Web Personalization Strategy

sociability (Preece, 2000). The intrinsic value of sociability is grounded in human motivation of being altruistic, interconnected, and seeking acceptance, attachment and care in interpersonal relationships (Argyle, 1996; Argyle & Lu, 1990). Positive social relationships give individuals a sense of well being and a feeling of comfort by creating support and a sense that they are not alone and are valued. Building a personalized network of social relationships is a self-fulfilling socio-emotional experience of satisfying the desire for social interaction, communication, and affiliation with others having similar interests or goals. There is mounting evidence of strong social ties built on social-emotional communications among participants of online communities (Fischer, Bristor, & Gainer, 1996; Granitz & Ward, 1996). The social personalization strategy enhances feelings of satisfaction and enjoyment by providing the social context important for fulfilling users’ needs for sociability. This suggests the following:

H6: The social personalization strategy will positively influence the perceived enjoyment of the Web site.

RESEARCH METHOD

The goal of this research is to develop and validate measures for different dimensions of personalization corresponding to the four distilled personalization archetypes. We first distilled the four personalization archetypes stated above from an extensive literature review of five general areas in which personalization has been studied: marketing/e-commerce, computer science/cognitive science, architecture/environmental psychology, information science, and social sciences, including sociology, anthropology and communication. Next, we generated and pretested 20 candidate measurement items for all four archetypes of personalization (see Table 2).

We used a lab experiment to test the validity and reliability of the proposed measurement model of personalization. We used the personalization features of actual commercial Web sites as the experimental stimuli. Because most Web sites tend to employ a combination of different personalization strategies, for example, Amazon.com implements both commercial (personalized recommendations) and social personalization (“Your Amazon friends” community strategies, the lab experiment allows us to measure participants’ responses to only one personalization feature of the Web site at a time. This helps to minimize the variations among participants’ responses due to exposure to mixed personalization features.

Experimental Design

A total of 229 undergraduate students (94 male, 135 female) enrolled in an introductory IS course were recruited to participate in the study for the benefit of extra course credit. Because this course is designed for non-IS majors, the sampling frame represents a wider range of disciplinary backgrounds than IS or business courses. Majors of subjects included chemistry, agriculture, liberal arts, and life sciences. This type of heterogeneity is advantageous in testing the reliability and generalizability of the measurement instrument (Shadish, Cook, & Campbell, 2001).

During the experiment, we showed the movie clips of someone using the personalization features of Web sites to the participants rather than having them actually visit the site. We did this for two reasons. First, having the participants view the movie clips ensures that they are only exposed to the personalization features that they are assigned to. Second, this approach reduces or eliminates the differences in participants’ responses resulting from irrelevant factors such as different browsing paths participants traversed.

To approximate actual Web browsing in the movie as closely as possible, we used Camtasia
Developing and Validating a Measure of Web Personalization Strategy

Table 2. Scale items for personalization strategy

<table>
<thead>
<tr>
<th>Construct</th>
<th>Code</th>
<th>Questionnaire Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architectural</td>
<td>ARCH1</td>
<td>Personalizing the site creates a delightful Web experience that is unique to me.</td>
</tr>
<tr>
<td></td>
<td>ARCH2</td>
<td>I am able to tailor the look and feel of the site to my own personal taste and style.</td>
</tr>
<tr>
<td></td>
<td>ARCH3</td>
<td>Personalizing the site creates a Web environment that is aesthetically pleasing to me.</td>
</tr>
<tr>
<td></td>
<td>ARCH4</td>
<td>I have a sense of control and mastery in creating my own space online.</td>
</tr>
<tr>
<td></td>
<td>ARCH5</td>
<td>It feels like decorating my own house when I’m personalizing this site.</td>
</tr>
<tr>
<td>Instrumental</td>
<td>INSTR1</td>
<td>Personalizing the site makes it more useful and usable for me.</td>
</tr>
<tr>
<td></td>
<td>INSTR2</td>
<td>I found personalizing the site helps me to obtain my goal for using the site more efficiently.</td>
</tr>
<tr>
<td></td>
<td>INSTR3</td>
<td>I would like to see other sites that I frequently view have the functionality this site provides.</td>
</tr>
<tr>
<td></td>
<td>INSTR4</td>
<td>Personalizing the site makes the interaction with the site easier for me.</td>
</tr>
<tr>
<td></td>
<td>INSTR5</td>
<td>The site provides many functions that I can configure to suit my own needs.</td>
</tr>
<tr>
<td></td>
<td>INSTR6</td>
<td>Personalizing the site helps me locate the right information/product/service I need.</td>
</tr>
<tr>
<td>Social</td>
<td>SOCIA1</td>
<td>Personalizing the site helps to connect me to a community that is potentially interesting to me.</td>
</tr>
<tr>
<td></td>
<td>SOCIA2</td>
<td>Personalizing the site helps to fulfill my needs for socialization and communication with others.</td>
</tr>
<tr>
<td></td>
<td>SOCIA3</td>
<td>Personalizing the site gives me a sense of community while maintaining my own self-identity.</td>
</tr>
<tr>
<td></td>
<td>SOCIA4</td>
<td>Personalizing the site helps to create a congenial social environment for me.</td>
</tr>
<tr>
<td>Commercial</td>
<td>COMM1</td>
<td>Personalizing the site enables faster checkout, transaction, or information scanning.</td>
</tr>
<tr>
<td></td>
<td>COMM2</td>
<td>My concern for privacy is a major factor in determining whether I would personalize the site.</td>
</tr>
<tr>
<td></td>
<td>COMM3</td>
<td>I’m made aware of new products or useful information regarding sales and promotion that I didn’t know before by personalizing the site.</td>
</tr>
<tr>
<td></td>
<td>COMM4</td>
<td>Personalizing the site helps the businesses know me better so that they can serve me better.</td>
</tr>
<tr>
<td></td>
<td>COMM5</td>
<td>I felt that this site is knowledgeable about my needs and wants in terms of what they can offer to me.</td>
</tr>
</tbody>
</table>

software to capture the entire course of Web interactions, dubbed with synchronized human voice to explain each mouse movement. To prime participants, we created four scenarios for each Web site describing the typical context and purpose of using the personalization features. It is important to make the usage context and purpose explicit to the subjects, as the conceptualization
and operationalization of the four personalization strategies are based on the user motivations and objectives underlying personalization. To ensure the best audio quality, we tested several volunteers and chose one with the most pleasant voice in speaking standard American English. A total of five AVI files were generated, including instructions for the participants, and one for each site. The resolution of the movie clip on screen is 1080x720.

To optimize the effect of movie viewing, we conducted the experiment in a state-of-the-art instructional lab, in which participants viewed the movie on a personal computer via a 19-inch flat screen LCD monitor using a headphone. For practical purposes, balanced incomplete block design (BIBD) was adopted for the experiment. To reduce judging fatigue and increase reliability of evaluations, each participant was randomly assigned to view only two of the four Web site stimuli, with order counterbalanced across the participants. Among the 458 data points collected from the experiment, 390 were useable after discarding the questionnaires with incomplete responses and inconsistent answers to the same question being asked twice.

**Experimental Web Sites**

In determining the Web personalization stimuli for the experiment, we chose to use actual Web sites rather than creating our own because actual ones are representative of the tradeoffs designers will have in creating them to meet actual business needs and therefore will enhance generalizability of the experimental results to other commercial Web site settings. We identified four prominent Web sites whose personalization strategies have addressed personal needs of a wide base of Web users and been advocated as exemplars of Web personalization in the literature (Amazon’s book recommendation system, Land’s End’s 3-D virtual model tool, MyMSN information portal, and Yahoo!Group). Then, we made an interpretive analysis of those personalization strategies to match them with the four personalization archetypes (see Table 3). The interpretive analysis began with an examination of those personalization strategies in relation to the various personal needs of Web users they are intended to fulfill. Making explicit the relations between personalization strategies and types of user needs allowed us to trace back to the motivations underlying these personalization strategies and identify a match between each personalization strategy and personalization archetype. In this way, the personalization archetypes were respectively operationalized through the selected Web personalization stimuli.

Amazon’s book recommendation system was used to operationalize commercial personalization. The virtual model tool on the Land’s End Web site was used to represent instrumental personalization. Architectural personalization was implemented as the customizable information portal on the MyMSN site, where users can personalize the content and look-and-feel of the Web page by choosing different themes and color palates. Yahoo!Group was used to represent social personalization. It is expected that each representative site’s personalization stimulus will score significantly different from the other three sites’ stimuli and score the highest on the personalization type it represents. In order to verify the sufficiency of manipulation, we followed the recommendation of using manipulation check in the experiment (Shadish et al., 2001). This would suggest the following criteria for checking the manipulation of the personalization stimuli:

**Criterion 1:** MyMSN personalization stimulus will score higher on architectural personalization than the other three sites.

**Criterion 2:** Amazon personalization stimulus will score higher on commercial personalization than the other three sites.
Table 3. Representative sites chosen in the study

<table>
<thead>
<tr>
<th>Personalization Archetype</th>
<th>Sample Web Site</th>
<th>Operationalization</th>
<th>Personalization Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architectural personalization</td>
<td>MyMSN portal Web page</td>
<td>Allowing users to: 1) customize Web content by selecting desired information to display; 2) change the layout of the Web page by moving modularized content block to desired position; and 3) design Web aesthetics by choosing desired theme and color palate.</td>
<td>Fulfilling user’s need for a personalized Web space that meets his/her information need and reflects his/her aesthetic taste and style.</td>
</tr>
<tr>
<td>Social personalization</td>
<td>Yahoo!Group</td>
<td>Allowing users to: 1) create and maintain newsgroups; 2) communicate through chat room and message board; and 3) share working space of files and documents.</td>
<td>Fulfilling user’s need for social interactions in ways compatible with his/her desired level of communality and privacy.</td>
</tr>
<tr>
<td>Instrumental personalization</td>
<td>Land’s End virtual model tool</td>
<td>Allowing users to: 1) build a customizable 3-D virtual model configured in one’s own body parameters; 2) examine instantaneous fitting effect by trying clothes on the model; and 3) view clothes from different angles.</td>
<td>Fulfilling user’s need for an online clothing shopping tool that enables user to “try clothes on” before purchasing.</td>
</tr>
<tr>
<td>Commercial personalization</td>
<td>Amazon book recommendation system</td>
<td>Recommending books to customers based on their past behavior, such as past purchases, searches, and browsing habit.</td>
<td>Fulfilling user’s need for highly personally relevant products.</td>
</tr>
</tbody>
</table>

**Criterion 3:** Land’s end personalization stimulus will score higher on instrumental personalization than the other three sites.

**Criterion 4:** Yahoo! Group personalization stimulus will score higher on social personalization than the other three sites.

**DATA ANALYSIS AND RESULTS**

Factor analyses were conducted to assess the reliability, and the discriminant and convergent validities of the personalization measures. We first performed exploratory factor analysis (EFA) on the pooled response data of all the four person-
alization stimuli, and derived four dimensions of personalization, which were consistent with our measurement model of personalization. Next, confirmatory factor analysis (CFA) was conducted on the response data of each personalization stimulus separately, to show that the measures of four dimensions of personalization and the three outcome constructs held relevant across the four sites, providing strong evidence of external validity for the measurement model. In addition, conducting separate analysis on four sites independently may provide site-specific knowledge on the relative importance of the constructs under investigation, given the different business natures and design features of these sites (Kim, Lee, Han, & Lee, 2002). Next, a series of planned contrast analyses are conducted for manipulation check purpose. Finally, we employed correlation and regression analyses to test the hypothesized relationships between personalization strategies and their cognitive and affective outcomes.

Results: Personalization Strategy Measures

We first computed reliability coefficients of the scales using Cronbach’s alpha on all 20 items. Based on each item’s contribution to reliability, we discarded two items (SOCIA1 and COMM2 and ARCH1) that showed very low corrected item-total correlations (0.44, 0.0004, and 0.29, respectively). Next, we removed two items (Instr5 and Instr3) that exhibited cross loading problems. After removing items with low reliability and cross-loading problems, the instrument consisted of 15 items, shown in the left-most column in Table 4.

Reliability

Table 3 presents the reliability coefficients in the right-most column, which are calculated from the EFA analysis of the data across four sites.

Table 4. Construct loading for personalization strategy instrument from exploratory factor analysis

<table>
<thead>
<tr>
<th>Scale Items</th>
<th>Instrumental</th>
<th>Social</th>
<th>Architectural</th>
<th>Commercial</th>
<th>Reliability Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSTR1</td>
<td>0.613</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INSTR2</td>
<td>0.886</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INSTR4</td>
<td>0.517</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INSTR6</td>
<td>0.520</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOCIA2</td>
<td></td>
<td>0.724</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOCIA3</td>
<td></td>
<td>0.828</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOCIA4</td>
<td></td>
<td>0.748</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARCH2</td>
<td></td>
<td></td>
<td>0.867</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARCH3</td>
<td></td>
<td></td>
<td>0.728</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARCH4</td>
<td></td>
<td></td>
<td>0.708</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARCH5</td>
<td></td>
<td></td>
<td>0.743</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMM1</td>
<td></td>
<td></td>
<td></td>
<td>0.559</td>
<td></td>
</tr>
<tr>
<td>COMM3</td>
<td></td>
<td></td>
<td></td>
<td>0.573</td>
<td></td>
</tr>
<tr>
<td>COMM4</td>
<td></td>
<td></td>
<td></td>
<td>0.660</td>
<td></td>
</tr>
<tr>
<td>COMM5</td>
<td></td>
<td></td>
<td></td>
<td>0.458</td>
<td></td>
</tr>
<tr>
<td>Eigenvalues</td>
<td>5.337</td>
<td>2.519</td>
<td>1.641</td>
<td>1.109</td>
<td></td>
</tr>
</tbody>
</table>
All coefficients are larger than 0.8, except for commercial personalization (0.6334).

Convergent Validity

As our theory predicted, four components had eigenvalues greater than 1. The four columns in the middle of Table 3 present the factor loadings for the four constructs of personalization from the EFA analysis based on data pooled from all four sites. All 15 items converge well on their corresponding constructs, with high loadings on the constructs they are intended to measure and low loadings on others (factor loadings that are less than 0.5 are not shown in the table).

Independent CFA analysis (see Table 5) on each site showed varied indices of goodness of fit for the measurement model. Among the four sites, Amazon seems to demonstrate the best fit (GFI=0.8582, CFI=0.9454, RMSEA=0.0588, chi-square/DF < 2). Yahoo!Group and MyMSN also obtained acceptable goodness of fit indices (Yahoo!Group: GFI=0.8604, CFI=0.8981, RMSEA=0.0602, chi-square/DF < 2; MyMSN: GFI=0.8272, CFI=0.8281, RMSEA=0.0882, chi-square/DF < 2), which supports the validity of the four measures for these three sites. Land’s End, on the other hand, did not have a good fit. One reason could be that the instrumental personalization features present in the experimental stimulus such as using the avatar and virtual product presence were not included in the measures of instrumental personalization, and thus were not properly measured by the current instrument. This result, however, provided us valuable feedback on how to improve the instrument.

Discriminant Validity

Table 6 presents the factor correlations and average variances extracted (AVE) for the four personalization constructs from the EFA analysis. The

<table>
<thead>
<tr>
<th>Websites</th>
<th>Confirmatory Factor Analysis</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Confirmatory Factor Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C h i</td>
<td>S q u a r e</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D F</td>
<td>GFI</td>
<td>RMSEA</td>
<td>CFI</td>
<td>NNI</td>
<td>NFI</td>
</tr>
<tr>
<td>Amazon</td>
<td>110.7545</td>
<td>0.8582</td>
<td>0.0588</td>
<td>0.9454</td>
<td>0.9317</td>
<td>0.8138</td>
</tr>
<tr>
<td>Land’s End</td>
<td>175.3926</td>
<td>0.7911</td>
<td>0.1112</td>
<td>0.7857</td>
<td>0.7322</td>
<td>0.6700</td>
</tr>
<tr>
<td>My MSN</td>
<td>140.8529</td>
<td>0.8272</td>
<td>0.0882</td>
<td>0.8281</td>
<td>0.7852</td>
<td>0.6768</td>
</tr>
<tr>
<td>Yahoo!Group</td>
<td>108.3229</td>
<td>0.8604</td>
<td>0.0602</td>
<td>0.8981</td>
<td>0.8726</td>
<td>0.6848</td>
</tr>
</tbody>
</table>

| Table 6. Factor correlation and AVE for personalization strategy constructs |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Instrumental | Social | Architectural | Commercial |
| Instrumental | 0.634 | 0.195 | -0.443 | 0.404 |
| Social | 0.7667 | 0.7615 | -0.141 | -0.207 |
| Architectural | 0.7615 | -0.133 | 0.5625 |
diagonal elements in Table 5 are the square root of AVE, and the off-diagonal elements are factor correlations. The four diagonal values are all larger than 0.5 and also larger than the correlation coefficients of their corresponding factors with other factors, which suggests that the measures have appropriate discriminant validity across all four sites (Gefen & Straub, 2000). The low factor correlations between pair-wise subscales of personalization also indicate the distinctive dimensionality of the four theoretical archetypes within the personalization construct.

**Results: Manipulation Check**

In order to check the representativeness of each personalization stimulus for its respective personalization strategy, a series of priori planned contrast tests were conducted to compare the performance of the four personalization stimuli based on their scores on different personalization strategies. Due to the nonorthogonality of treatment and individual effects in this study (incomplete block design), we performed a combined intra- and inter-block analysis using SAS Proc Mixed procedure to estimate the treatment least-squares means and treatment contrasts. Table 8 shows the contrast coefficients, the personalization type as dependent variable, the t value, and the significance level for each contrast test.

Contrast 1 (testing criteria 1) compared the commercial personalization scores between the Amazon personalization stimulus and the other three Web sites (Land’s End, My MSN, and

**Figure 1. Site comparison chart**

![Figure 1. Site comparison chart](image-url)
Developing and Validating a Measure of Web Personalization Strategy

Yahoo!Group): Contrast 2 (testing criteria 2) tested whether the instrumental personalization score for the Land’s End stimulus is higher than the other three Web sites’ scores (Amazon, My MSN, and Yahoo!Group). In contrast 3 (testing criteria 3), we tested whether the My MSN stimulus scores higher than the other three Web sites (Amazon, Land’s End, and Yahoo!Group) on architectural personalization. Finally, in contrast 4 (testing criteria 4), we compared the social personalization score for the Yahoo!Group stimulus with those scores of the other three Web sites (Amazon, Land’s End, MyMSN).

As expected, the results indicate that the Amazon stimulus scored significantly higher than the other three Web sites on commercial personalization; the architectural personalization score for the My MSN stimulus was significantly higher than those for the other three Web sites; and the Yahoo!Group stimulus had a significantly higher score on social personalization than the other three Web sites. However, the Land’s End stimulus did not differ from the other Web sites on the score of instrumental personalization. Therefore, criteria 1, 3, and 4 were supported by the planned contrast tests, but criteria 2 was not met. An examination of Figure 1, which visualizes the performance of each personalization stimulus on each of the four personalization archetypes, reveals that all personalization stimuli except Land’s End personalization stimulus scored significant higher on two personalization archetypes: the personalization archetype they each represent and instrumental personalization. The reason may be that all personalization strategies are to some extent instrumental in a sense that they all provide a means to fulfill individual’s personal needs or goals, be it creating an aesthetically appealing interface, building a personal social network, or purchasing the products satisfying an individual’s needs. Although Land’s End personalization stimulus did not score higher on instrumental personalization than the other personalization stimuli, it still had a significantly higher score on instrumental personalization than on any of the other three personalization archetypes, as shown by the results of pairwise comparisons between its scores on instrumental personalization and on each of the other three personalization archetypes (all three P values < 0.001 after Bonferroni adjustments). Hence, this confirmed our expectation that Land’s End personalization stimulus would be perceived to be more instrumental than architectural, social, or commercial.

RESULTS: ASSOCIATION BETWEEN PERSONALIZATION STRATEGY AND OUTCOME

In order to test H1 through H6, we conducted three multiple regressions, each using mean personalization score as the independent variables and one of the outcome scores as the dependent variable. Table 7 shows the results of the multiple regressions, in which the scores of perceived usefulness, perceived ease of use, and enjoyment of Web site were respectively regressed on the mean scores of four personalization dimensions. All four multiple regression analyses were significant. We discuss the results related to the hypotheses testing as follows.

Hypotheses 1 and 2 suggest that instrumental and commercial personalization will influence perception of Web site usefulness. The results of multiple regression 1 show that social personalization (Beta = 0.336, t = 5.562, p < 0.001) and instrumental personalization (Beta = 0.396, t = 5.916, p < 0.001) strategies are strong and significant predictors of perceived usefulness of the Web site. However, the effect of commercial personalization was not significant. So, H1 was supported, but H2 was rejected.

Hypotheses 3 and 5 were supported by the results of the second multiple regression, which show that both instrumental personalization (Beta = 0.279, t = 3.696, p < 0.001) and architectural...
Table 7. Association between personalization strategy and outcome

<table>
<thead>
<tr>
<th></th>
<th>Usefulness</th>
<th>Ease of Use</th>
<th>Enjoyment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Beta</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>0.061</td>
<td>0.116</td>
<td>0.168*</td>
</tr>
<tr>
<td>Instrumental</td>
<td>0.396**</td>
<td>0.279**</td>
<td>0.259**</td>
</tr>
<tr>
<td>Architectural</td>
<td>0.039</td>
<td>0.194*</td>
<td>0.425**</td>
</tr>
<tr>
<td>Social</td>
<td>0.336**</td>
<td>0.063</td>
<td>0.064</td>
</tr>
</tbody>
</table>

* P value < 0.05  
** P value < 0.001

personalization (Beta = 0.194, t = 2.727, p = 0.007) strategies have significant, positive effects on the user’s perceived ease of use of the Web site.

Consistent with hypothesis 4, the results of the third multiple regression indicate that implementing the architectural personalization (Beta = 0.425, t = 7.164, p < 0.001) strategy of a Web site has strong, significant effect on users’ feeling of enjoyment when they interact with the Web site. We also found that instrumental (Beta = 0.259, t = 4.109, p < 0.001) and commercial personalization (Beta = 0.168, t = 2.866, p = 0.005) strategies can influence users’ perceived enjoyment of using the Web site. However, hypothesis 6 was rejected because the effect of social personalization on perceived enjoyment of the Web site was not significant.

**DISCUSSION**

The purpose of this study was to develop and validate measurement scales for personalization by identifying four distinctive personalization archetypes and hypothesizing their respective relationships with different cognitive and affective outcomes of usage. This effort yielded promising results in several respects. New scales for measuring personalization strategies were developed based on the definitions of personalization archetypes, which were derived from the review of extensive multidisciplinary studies. A total of 25 items initially generated based on the four archetypes was honed down to 20 items, five for each personalization type. A lab experiment with 229 student subjects was conducted to explore the structure of the personalization construct and to validate the instrument of using four Web site stimuli, respectively representing four types of personalization strategies. According to the results of data analysis, the measurement scales of four dimensions of personalization exhibited sufficient psychometric properties in terms of reliability and validity. Convergent and discriminant validities were supported by both the analysis of pooled data across all four sites and the independent analysis.
Developing and Validating a Measure of Web Personalization Strategy

on each site separately. The study empirically supported our theoretical model of the structure of personalization construct, which consists of four distinct dimensions: architectural, instrumental, social and commercial.

Because of the multidimensional nature of the personalization construct, we should not use only a single yardstick to measure the effectiveness of personalization strategies. We identified two major categories of use outcomes associated with extrinsic motivation and extrinsic motivation, respectively. To test the relationship between personalization and extrinsic motivation, we hypothesized that instrumental and commercial personalization would positively influence two TAM constructs: perceived usefulness and ease of use. To test the relationship between personalization and intrinsic motivation, we hypothesized that social and architectural personalization would positively influence perceived enjoyment and ease of use. Our hypotheses were partially supported by the empirical results which revealed additional interesting findings.

Among significant findings, instrumental personalization was found to have a positive influence not only on perceived usefulness and ease of use but also on perceived enjoyment. Instrumental personalization facilitates perceived ease of use as it is driven by the design philosophy that emphasizes clarity, consistency, freedom from ambiguity and error (Norman, 1988). Perceived ease of use concerns the degree of “free of effort” (Davis, 1989), and low attention allocation to the system interaction, whereas enjoyment relates to the highly aroused emotions and focal attention allocation. With the presence of ease of use, enjoyment can be facilitated by the availability of additional attention resources. The absence of ease of use may hamper enjoyment because focal attention has to be allocated to “cope” with the system. This leads to our speculation that ease of use may be the antecedent to enjoyment. Note that instrumental personalization is the only strategy that lands itself on all three use outcomes. To some degree, the result adheres to the TAM prediction that ease of use mediates usefulness.

Contrary to our hypothesis, we did not find significant linkage between social personalization and enjoyment. Instead, the results suggest that social personalization is associated with perceived usefulness. We offer two explanations for this result. First, viewing the movie clip instead of experiencing the actual social interaction via the personalized communication platform may have mitigated the social-emotional impact of the stimulus on participants. The stimulus may

| Contrast | Personalization Type | Website |  | Land’s End | My MSN | Yahoo!Group | T Value |
|----------|----------------------|---------|  |           |       |             |        |
| 1        | Commercial           | 3       |  | -1        | -1    | -1          | 6.40** |
| 2        | Instrumental         | -1      |  | 3         | -1    | -1          | 0.73   |
| 3        | Architectural        | -1      |  | -1        | 3     | -1          | 11.81**|
| 4        | Social               | -1      |  | -1        | -1    | 3           | 17.53**|

* P value < 0.05
** P value < 0.001
have simply been too weak to elicit the expected response. Although we are confident in the ability of the movie to simulate the visual and audio qualities of browsing, as well as the level of detail of the Web interaction process captured in the movie, viewing interaction taking place may not be a perfect surrogate for actually being in the interaction. Secondly, the result reveals that social personalization serves some useful goals external to the socializing process. People engage in social personalization not just for pure experience only, but for a purpose too. This would be consistent with the view of relationship as social capital (Wellman, 2002).

Finally, contrary to our expectation, we did not find significant association of commercial personalization with perceived usefulness, but rather a small association with enjoyment (0.168). This might be due to the following reasons. On the one hand, commercial personalization usually requires users to set up personal accounts in order to facilitate future transactions, which can be tedious and time-consuming and hence may not lead to perceived usefulness of the system. On the other hand, providing personally relevant products to the customers through commercial personalization can lead to high levels of satisfaction and evoke a feeling of enjoyment in the customers.

**CONCLUSION: THEORETICAL AND PRACTICAL CONTRIBUTIONS**

In sum, the main contribution of this research is two-fold: theoretical and practical. Theoretically, the study answers the call for conceptualizing the multidimensional personalization construct. As previous empirical studies have shown, a one-dimensional scale was not sufficient to tap into the different effects yielded by personalization construct. We have identified four dimensions underlying the personalization construct and developed a measurement scale representing the four dimensions. The instrument exhibited sufficient psychometric properties on the calibrated Web site that are representative of respective dimensions.

In practice, this instrument can be utilized in at least two ways. Firstly, the instrument we developed in this research can be used as a guideline for developing Web personalization strategies because it provided four basic means to personalize Web sites; that is, by providing functionalities and information specifically needed by the user so as to enhance efficiency and productivity, by tailoring the interface to the user’s own taste and style, and by enabling interaction and connectivity specifically for the user’s social network. For productivity-oriented personalization systems, key usability issues to consider would be ease of use, clarity, consistency, free from ambiguity, and error. The aspect of ease of use includes both the use of the application itself and the setup and configuration to make personalized features functional. Consistency helps users better orient themselves to the site and alleviates cognitive effort. The enjoyment or entertainment-oriented personalization applications capitalize on the process and experience of using the systems. They are designed to stimulate thinking and to invoke feelings. The results are not tangible, but the process itself is critical in creating an engaging, fulfilling user experience. The principle of consistency may not be sufficient to invoke feelings or engage users on the site for an extended amount of time. In addition, the instrument can measure users’ perception of different personalization strategies and be used as a criterion for evaluating the effectiveness of the implementation of personalization strategy. For example, at the design and testing stage, Web designers and researchers can evaluate the performance of the site in terms of its ability to cater to users’ personal needs by having the users rate the site using the instrument. Weak scores would indicate potential areas for improvement.
Secondly, we empirically tested the hypothesis that different personalization strategies lead to different cognitive and affective outcomes of usage. As extant literature shows, measuring the effectiveness of Web site personalization using a monolithic method such as Return-On-Investment or click-to-buy ratio is not sufficient to gauge the multidimensional nature of personalization. As Web users come to interact with the site with different motivations, which largely dictates their usage expectation and online behavior (Davis et al., 1992; Venkatesh, 2000), understanding the underlying motivations for using Web personalization features and how different motivations relate to respective cognitive and affective outcome is crucial for realizing the potential of Web personalization. This study is our initial attempt along this line of research. Drawing on the existing literature on the distinction between extrinsic and intrinsic motivations, we empirically tested different usage outcomes by using different Web personalization features. Specifically, we found significant and strong correlations between architectural personalization and perceived enjoyment and ease of use, and between instrumental personalization and perceived usefulness and ease of use. The empirical data also suggested significant, small-sized correlation between social personalization and perceived usefulness, and between commercial personalization and perceived enjoyment. Our initial hypotheses were partially supported and new insights were obtained.

LIMITATIONS

Several limitations should be considered when interpreting the results of this study. First, the data were collected from a convenience sample of students, which may restrict the generalizability of the results. Although college students are representative of many young Web users and online consumers in the real world, the participants of this study are heavier Internet users and more frequent online shoppers than other segments of the general population. Their perceptions of personalization may be different from other segments.

Second, having the participants watch the movie clips instead of letting them interact with the Web sites freely has some drawbacks. Although this method allowed us to have a tighter control over the experiment (e.g., to ensure that the participants are only exposed to the personalization features that they are assigned to, and to reduce or eliminate the differences in participants’ responses resulting from irrelevant factors such as different browsing paths participants traversed), watching other people using the Web personalization features may be different from experiencing the features by oneself. This might introduce error into the data because the environment of the experiment may not be ecologically representative. In addition, because the conceptualization and operationalization of the four personalization strategies are based on the user motivations and objectives underlying personalization, we also provided audio instructions to make the usage context and purpose of the experimental personalization features explicit to the participants. However, priming the participants to a particular personalization regimen could produce bias in the participants’ responses.

Third, the positivist research approach of this study introduced further limitations. Positivist research emphasizes objectivity, which separates the researcher from the respondent. It is practiced through standardized process, control of possible influential variables, and pre-determined structured questions that facilitate measurement. This approach leads respondents to reply in a restricted range of ways rather than let them answer the questions and elaborate on their replies in their own ways. As a result, the positivist research may fail to yield an authentic understanding of the respondents’ opinions, attitudes, feelings, and experiences. Therefore, the results of this study
Developing and Validating a Measure of Web Personalization Strategy

may not represent a full insight into the participants’ view of Web personalization features.

Finally, another limitation of this study is concerned with the four criteria we developed to check how well each stimulus represents its corresponding personalization archetype. Because the focus of this study is to differentiate the four personalization archetypes and their respective cognitive and affective effects, the criteria examine each stimulus’ representativeness of its corresponding personalization archetype relative to the other three stimuli. Therefore, a criterion will be supported if the stimulus of focus is shown to represent its corresponding personalization archetype better than the other three (i.e., the stimulus has a higher score on its corresponding personalization archetype than the other three). However, meeting such a criterion may not be sufficient to establish the stimulus of focus as a successful operationalization of its corresponding personalization archetype, because the criterion evaluates the performance of stimulus using its relative score rather than its absolute score on the corresponding personalization archetype.

Despite the abovementioned limitations of this study, we do believe our results provide valuable insights on different dimensions of personalization and their influences on users’ cognitive and affective responses. This research is the beginning of a rich stream of research investigating the multidimensional nature of personalization. We call for further studies along this line of research using the validated instrument of personalization, and testing it in various settings.

FUTURE DIRECTIONS

There are several possible ways of continuing this research. First, the current study suggested ways of improving the instrument. Revising the items for the instrumental personalization by focusing on the design aspects of instrumental personalization, such as providing, enabling and delivering useful personalization tools, is our next step toward a more valid and stable instrument. The improved instrument can be used for: 1) stimulus scaling, that is, to scale Web sites of different domains (online retailer, sports, information portal, etc.) and industries (finance, travel, entertainment, education, government, etc.). The purpose is to discover patterns of Web personalization strategies within and across domain or industry, which can be used to inform further improvement. 2) Subject scaling, that is, to scale Web users along the dimensions of personalization so that a personal profile based on individual’s response will be created. It would be another tool for business to further understand their users.

In conclusion, the overarching goal in this article was to enrich our understanding of the multidimensionality of the personalization construct, and how different dimensions of personalization elicit different cognitive and affective user responses. Given that personalization is becoming an important offering of IT artifact, such research has value for theory development as well as for practice. Several avenues for further study are suggested in hope of a fruitful buildup of a continuous body of personalization research.

REFERENCES


Developing and Validating a Measure of Web Personalization Strategy

The DATA BASE for Advances in Information Systems, 28(2), 53-62.


Chapter 2.29
User Interface Formalization in Visual Data Mining

Tiziana Catarci
University of Rome “La Sapienza”, Italy

Stephen Kimani
University of Rome “La Sapienza”, Italy

Stefano Lodi
University of Bologna, Italy

ABSTRACT
Despite the existence of various data mining efforts that deal with user interface aspects, very few provide a formal specification of the syntax of the interface and the corresponding semantics. A formal specification facilitates the description of the system properties without being concerned about implementation details and enables the detection of fundamental design issues before they manifest themselves in the implementation. In visual data mining, a formal specification can enable users to decide which interaction/operation to apply to get a desired result; help users to predict the results of their interactions/operations with the system; and enable the development of a general interaction model that designers/developers can use to understand the relationships between user interactions and their compositions. In this work, we describe an approach for formalizing the visual interface of a core data mining system, which has been employed in the development of a visual data mining system named VidaMine.

INTRODUCTION
In this day and age, data still present formidable challenges to effective and efficient discovery of knowledge. It should be acknowledged that a lot of research work has been and is being done with respect to knowledge discovery (KD). Much of the work has concentrated on the development and the optimization of data mining algorithms using techniques from other fields such as artificial intelligence, statistics, and high performance computing (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b). Besides various glaring issues (such as
the need to have an overall framework that can support the entire discovery process, supporting human involvement in the entire process, etc. Mannila observes that relatively little research work has been published on the theoretical foundations of data mining (Mannila, 2000). On the same note, although there are many data mining efforts that deal with user interface issues, very few efforts provide or give a precise definition of the syntax of the user interface and the corresponding semantics.

Formal specifications enable the description of the system properties without having to be concerned about implementation details. The system properties are often specified using a precise notation. The specification can be used to construct models of the system. Formal methods make the analyst think abstractly about the problem at hand and the corresponding system thereby exposing fundamental design decisions well in advance before they manifest themselves in the implementation of the system. While it is true that the formal specification should not determine the programming aspects (e.g., algorithms and data structures), it should describe the behavior of the system in a precise or rigorous manner. Moreover, with a formal specification it is possible to transform a system model while preserving important properties of the model. In practice, a formal approach in which each and every design decision is proven to be a correct refinement step is rarely performed due to the high costs involved. However, substantial application of refinement does considerably improve the understanding of the design process. A formal specification of the visual interface of a data mining system can facilitate the gathering of information about the most useful usage patterns, which can then be used to guide the design and layout of user interfaces for visual data mining. Moreover, a formal specification of the visual interface can facilitate automated (and/or objective) evaluation of the usability of the user interface of a visual data mining system.

In information visualization, various specifications/models for characterizing visualization aspects have been proposed such as (Baudel, 2004; Chi & Riedl, 1998; Chuah & Roth, 1996). In fact, it does turn out that most of the efforts that are related to our work are mainly found in information visualization and exploration efforts rather than in core data mining. Some of the benefits of specifications/models such as the foregoing do apply to visual data mining as well, where visualization tends to be a key ingredient. Consequently and borrowing from Chi et al. (1998), a similar formal specification in visual data mining can: enable users to decide which user interaction/operation to apply in order to get a desired result; help users to predict the results of their interactions/operations with the visual data mining system; and enable the development of a general interaction model that designers/developers can use to classify and understand the relationships between user interactions and the composition of interactions. In fact, such a model could help eliminate errors caused by other imprecise or incorrect models. In this work, we describe an approach for formalizing the visual interface of a core data mining system. The proposed approach has been employed in the development of a visual data mining system named VidaMine.

**BACKGROUND**

**Knowledge Discovery**

Knowledge discovery (KD) may be defined as the process of identifying valid, novel, potentially useful, and ultimately understandable models and/or patterns in data (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996a; Fayyad et al., 1996b). On the whole, the knowledge discovery process may be defined as an interactive and iterative non-trivial process that entails various phases as seen in Figure 1.
The KD phases include carrying out some initial planning (understanding the application domain, relevant prior knowledge, and goal/s of the user), data integration, selection of target data, data cleaning and pre-processing, data reduction and transformation, selection of suitable data mining techniques to support the discovery process, and evaluation, presentation, and interpretation of results. Through carrying out the phases, the KD process intends to find a subset of results that may be considered as new “knowledge” (Fayyad et al., 1996a, 1996b). KD is of interest to researchers in many research disciplines such as machine learning, pattern recognition, databases, statistics, artificial intelligence, expert systems, and information visualization.

**Data Mining**

Data mining is a core step in the knowledge discovery process that, under acceptable computational efficiency limitations, enumerates models and patterns over the data (Fayyad et al., 1996a, 1996b). It is worth recalling that the knowledge discovery process is as an interactive and iterative non-trivial process that entails: carrying out some initial planning (understanding the application domain, relevant prior knowledge, and goal/s of the user), data integration, selection of target data, data cleaning and pre-processing, data reduction and transformation, selection of suitable data mining techniques to support the discovery process, and evaluation, presentation, and interpretation of results from which there exists a subset that may be considered as new “knowledge” (Fayyad et al., 1996a, 1996b). Data mining methods include clustering, classification, regression, characterization, dependency modeling change and deviation detection, and pattern-based similarity matching.

The primary goals of data mining are verification and discovery. The verification goal aims at validating some hypotheses based on specific user needs. The user generates a series of hypothetical patterns and relationships (assumptions or claims). The user then formulates and issues
queries to the system (actually to the data itself) to verify (or disprove) the claims. The discovery goal involves finding “new” patterns or discovering new knowledge. Rather than verify hypothetical patterns, the goal here is to use the data itself to uncover or identify such patterns. Such methods of discovery may be initiated based on the guidance of a user to analyze a certain domain through a predetermined perspective or by automated learning. Discovery can be predictive or descriptive. Prediction entails “foretelling” unknown or future values of the same variables or other variables of interest whereas description involves getting an interpretation/understanding of the data. Classification, regression, and time series models are primarily useful for prediction. Clustering, association, and sequence discovery models are primarily useful for description of the behavior that is captured in the data.

Rules

Metaqueries

Metaquerying (Mitbander, Ong, Shen, & Zaniolo, 1996) is a data mining technique that is especially useful in mining relational and deductive databases. Metaqueries (or metapatterns) provide a generic description of a class of patterns that the user may want to discover from the underlying dataset. With metaqueries, it is possible to mine patterns that link several tables in the target dataset. Metaquery specification can be carried out manually (for instance by an expert user). Alternatively, the specification can be automated by exploiting the schema of the underlying dataset.

Let \( U \) be a countable domain of constants. A database \( DB \) is \( (D, R_1, ..., R_n) \) where \( D \subseteq U \) is finite, and each \( R_i \) is a relation of fixed arity \( a(R_i) \) such that \( R_i \subseteq D^{a(R_i)} \).

A metaquery is a second-order template of the form (Angiulli, Ben-Eliyahu-Zohary, Ianni, & Palopoli, 2000):

\[ T \leftarrow L_1, ..., L_m \]

where \( T \) and \( L_i \) are literal schemes. Each literal scheme \( T \) or \( L_i \) is of the form \( Q(Y_1, ..., Y_n) \) where \( Q \) is either a predicate (second-order) variable or a relation symbol, and each \( Y_j (1 \leq j \leq n) \) is an ordinary (first-order) variable. If \( Q \) is a predicate variable, then \( Q(Y_1, ..., Y_n) \) is called a relation pattern of arity \( n \), otherwise it is called an atom of arity \( n \).

The left-hand side \( T \) is called the consequent or the head of the metaquery. The right-hand side \( L_1, ..., L_m \) is called the antecedent or the body of the metaquery. Consider the relations CustCent, ClustOutI and ServCent with the following attributes: CustCent.CustID, CustCent.CentID, ClustOutI.CustID, ClustOutI.ServID, ServCent.ServID and ServCent.CentID. The following is an example of a corresponding metaquery:

\[ CustCent(CustID, CentID) \leftarrow \{ClustOutI (CustID, ServID), ServCent(ServID, CentID)\} \]

Intuitively, given a database instance \( DB \), answering a metaquery \( MQ \) on \( DB \) amounts to finding all substitutions \( \sigma \) of relation patterns appearing in \( MQ \) by atoms having as predicate names relations in \( DB \) such that the Horn rule \( \sigma(MQ) \) (which is obtained by applying \( \sigma \) to \( MQ \)) encodes a dependency between the atoms in its head and body. The Horn rule is supposed to hold in \( DB \) with a certain level of plausibility/relevance. The level of plausibility is based on measures of interestingness such as support and confidence. The measures of support and confidence are described in Section “Support and Confidence.”

Metaqueriers have been applied in the telecommunication industry, in a common-sense knowledge base, and in the chemical industry (Leng & Shen, 1996). Metaqueries have also been applied in analyzing time sequence data for semiconductor...
ASSOCIATION RULES

Association rules were introduced in Agrawal, Imielinski, and Swami (1993). Association rules represent a data mining technique that is used to discover implications between sets of items in the database.

Let \( I = I_1, I_2, ..., I_m \) be a set of data items or literals and \( D \) a set (or database) of transactions, in which each transaction \( T \) is a set of items from \( I \) (i.e. \( T \subseteq I \)). Each transaction \( (T) \) is assigned some unique identifier, \( TID \).

Let \( X \subseteq I \) and \( Y \subseteq I \). A transaction \( T \) is said to contain \( X \) if \( X \subseteq T \). An association rule is an implication of the form:

**Equation 3**

\[
Y \rightarrow X
\]

where \( X \cap Y = \emptyset \). The left-hand side, \( Y \), is the consequent or the head of the association rule whereas the right-hand side, \( X \), is the antecedent or the body of the association rule.

The problem of mining association rules is to generate all association rules with a degree of relevance/interestingness that is greater than a certain minimum (such as user-specified) value. The problem of discovering all association rules can be decomposed into two sub-problems (Agrawal et al., 1993):

1. Finding all sets of items (itemsets) that have support above the minimum support. The measure of support is described in Section “Support and Confidence.” Itemsets with minimum support are called large itemsets and all others small itemsets.
2. Using the large itemsets to generate the desired rules. The desired rules are those whose confidence is greater or equal to the minimum confidence. The measure of confidence is described in Section “Support and Confidence.”

Computing the solution to the first sub-problem naively is very expensive, and efficient techniques to reduce its complexity have been researched extensively (e.g., Agrawal et al., 1993; Agrawal & Srikant, 1994). Solving the second sub-problem is comparatively much easier, since it amounts to computing, for every frequent itemset, the confidence of the association rules obtained by picking one item as the consequent of the rule from the frequent itemset.

Association rules are valuable and readily applicable in many areas, including marketing, catalog design, business management, decision-making, add-on sales, mailing, customer segmentation, and store layout. For instance, a superstore dealing in clothing may find that customers who buy ties also tend to buy shirts at the same time (same “market basket”). Consequently, the following would be an association rule with fairly high measures of interestingness:

**Equation 4**

"shirt" \( \rightarrow \) "tie"

SUPPORT AND CONFIDENCE

The relevance of a rule may be determined by measures of interestingness. In general, measures of interestingness measure the overall value of a rule with respect to utility, novelty, certainty, and simplicity. In the mining of metaqueries and association rules, there are two common measures namely confidence and support.

Let \( D \) be the target dataset. The support of a rule in \( D \) is the fraction or percentage of tuples in \( D \) that contain the union of the antecedent and the consequent. For instance, the association rule

\[
Y \rightarrow X
\]

where \( X \cap Y = \emptyset \), is said to contain \( X \) if \( X \subseteq T \). An association rule is an implication of the form:

**Equation 3**

\[
Y \rightarrow X
\]
in Equation 3 has support $supp$ in $D$ if $supp\%$ of transactions in $D$ contain $X \cup Y$. The confidence of a rule in $D$ is the fraction or percentage of tuples in $D$ containing the antecedent, that also contain the consequent. For instance, the association rule in Equation 3 has confidence $conf$ in $D$ if $conf\%$ of transactions in $D$ that contain $X$ also contain $Y$. Intuitively, support indicates how frequently the items in the rule occur together in the transactions of the database, and therefore represents the utility of the rule, whereas confidence indicates the strength of the implication represented by the rule.

## Clustering

Clustering is a process through which the target dataset is divided into groups of similar objects, such that the objects in a particular group are dissimilar to objects in other groups. Each such group is referred to as a cluster. Clustering is applicable in many arenas such as in analyzing astronomical data, in demographics, in insurance, urban planning, and Web applications.

### Classification of Clustering Methods

Traditionally, clustering methods have been classified into a taxonomy having two broad groups: hierarchical and partitional (Jain, Murty, & Flynn, 1999).

#### Hierarchical Clustering

Hierarchical methods produce a sequence of nested partitions. A compact way to represent nested partitions is by a dendrogram, i.e., a tree having single objects as leaves, showing the hierarchical relationships among the clusters. It is therefore possible to explore the underlying dataset at various levels of granularity. Hierarchical methods are further subdivided into agglomerative and divisive (Jain & Dubes, 1988; Kaufman & Rousseeuw, 1990).

Agglomerative (bottom-up) clustering starts with singletons, i.e., with each cluster containing exactly one point. The clustering then recursively merges two or more most appropriate clusters. The process goes on until a stopping criterion is fulfilled (such as the number of clusters input by the user). Examples of agglomerative algorithms include CURE (clustering using representatives) (Guha, Rastogi, & Shim, 1998), and CHAMELEON (Karypis, Han, & Kumar, 1999).

Divisive (top-down) clustering starts with one single cluster of all objects, and recursively subdivides the most appropriate cluster. The process goes on until some criterion is met. The PDDP (principal direction divisive partitioning) (Boley, 1998) algorithm is an example of divisive algorithms. Also in this category of divisive clustering are approaches based on the k-means algorithm (Hartigan, 1975; Hartigan & Wong, 1979) such as the bisecting k-means algorithm (Steinbach, Karypis, & Kumar, 2000; Wang, Wiederhold, Firschein, & Wei, 1998).

#### Partitional Clustering

Partitional methods attempt to identify clusters directly either by iteratively relocating points between subsets, or by associating clusters with the areas that are densely populated with data. Consequently, partitional methods fall into two categories: relocation methods and density-based methods.

Relocation methods focus on how well points fit into their clusters. Such methods intend to ensure that the built clusters have the proper shapes. Relocation methods are further subdivided into probabilistic, k-medoids, and k-means. The probabilistic clustering model is based on the assumption that data has been independently drawn from a mixture model of several probability distributions. The results of probabilistic clustering are often easy to interpret. Probabilistic clustering algorithms include SNOB (Wallace & Dowe, 1994), AUTOCLASS (Cheeseman &
In clustering methods that adopt the k-medoids approach, a cluster is represented by one of its points. When the medoids are selected, clusters are considered to be subsets of points close to respective medoids. Algorithms based on the k-medoid approach include PAM (partitioning around medoids) and the algorithm CLARA (clustering large applications) (Kaufman et al., 1990), CLARANS (clustering large applications based upon randomized search) (Ng & Han, 1994). In k-means (Hartigan, 1975; Hartigan et al., 1979), a cluster is represented by its centroid, which is a componentwise arithmetic mean of points within a cluster. Although the k-means algorithm does not work well with a dataset that has categorical attributes, the algorithm is an appropriate choice for datasets with numerical attributes.

Density-based methods aim at identifying connected components/areas in the dataset that are dense with data. In this respect, a cluster therefore corresponds to a connected dense component. Density-based methods can be further divided into two main categories: density-based connectivity and density functions. Density-based connectivity approach reduces density to a training data point. Algorithms that use the approach include DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), OPTICS (ordering points to identify the clustering structure) (Ankerst, Breunig, Kriegel, & Sander, 1999), DBCLASD (distribution based clustering of large spatial databases) (Xu, Ester, Kriegel, & Sander, 1998). Density functions approach reduces density to a point in the attribute space. DENCLUE (Hinneburg & Keim, 1998) is an example of an algorithm based on density functions. In fact, DENCLUE is a blend of density-based clustering and grid-based preprocessing.

Miscellaneous

There exist many other clustering techniques that do not fit well in one of the foregoing categories. For instance grid-based techniques, co-occurrence techniques, etc.

Grid-based techniques work indirectly with data by constructing summaries of data over the attribute space subsets. They segment the space and then aggregate appropriate segments. On the one hand, grid-based methods often use hierarchical agglomeration as a phase in their processing. Algorithms that use this approach include BANG (Schikuta & Erhart, 1997), STING (statistical information grid-based method) (Wang, Yang, & Muntz, 1997), and WaveCluster (Sheikholeslami, Chatterjee, & Zhang, 1998). On the other hand, the idea behind grid-based methods is exploited by other types of clustering algorithms (such as CLIQUE (clustering in quest) (Agrawal, Gehrke, Gunopulos, & Raghavan, 1998), MAFIA (merging of adaptive finite intervals) (Goil, Nagesh, & Choudhary, 1999; Nagesh, Goil, & Choudhary, 2001) as an intermediate phase in their processing.

Co-occurrence techniques are meant to handle special requirements when it comes to clustering categorical data. Algorithms ROCK (Guha, Rastogi, & Shim, 1999), SNN (shared nearest neighbors) (Ertöz, Steinbach, & Kumar, 2003), and CACTUS (clustering categorical data using summaries) (Ganti, Gehrke, & Ramakrishnan, 1999).

The Proposed Clustering Taxonomy and Framework

The foregoing traditional categorization of clustering methods into two broad groups—hierarchical and partitional (Jain et al., 1999)—is technically sound and relevant to various application domains. However, such categorization does not highlight similarities and differences between the various definitions of a cluster that are implicit in the methods. For instance, Ward’s minimum-variance method (Ward, 1963) and the PAM method PAM (partitioning around medoids) (Kaufman et al., 1990) are similar. However, the former is hierarchical whereas the latter is partitional.
As an alternative to the foregoing traditional approach of categorizing clustering methods, clustering can be regarded as an optimization problem, in which the function to be optimized is a mathematical measure of homogeneity or separation (Hansen & Jaumard, 1997). Such a perspective enables one to categorize clustering methods according to a taxonomy of homogeneity or separation functions. Therefore, such a perspective provides recourse for categorizing clustering methods.

Moreover, such taxonomy expresses cluster definitions in an implicit manner. Such categorization is most likely more effective in capturing different behaviors in practice. It therefore provides a more natural avenue for the process of selecting a clustering algorithm, which is most suited to a particular application or domain.

In this research work, we therefore adopt this perspective and to the best of the research survey, this is actually the first effort that uses a uniform framework for selecting clustering algorithms. More details about the same and how it has been applied in our design and implementation efforts will be discussed in Section “Formal Specification of the Visual Interface.”

**USER INTERFACE FORMALIZATION IN VISUAL DATA MINING**

**Introduction to VidaMine**

VidaMine is a visual data mining system that exploits various visual strategies thereby offering a consistent, uniform, flexible visual interface that allows or enables the user not only to process data, but also to steer, guide, or direct the entire process of data mining (DM) (Kimani, 2002). We adopted a user-centered user interface design,
equipped with usability studies. As reported in Kimani, Catarci, & Santucci (2003), we employed various usability methods progressively in the development lifecycle. The visual interface of VidaMine offers visual interaction environments across different mining techniques and tasks. At present, the system offers visual environments for mining metaqueries, performing clustering, and mining association rules. Figure 2, which is for the metaquery environment, is an illustration that shows the overall outlook and feel of the visual interface.

Formal Specification of the Visual Interface

In developing a formal specification for the visual interface, we propose the provision of two specifications: an abstract formal specification that is defined for each of the mining methods currently supported by VidaMine and a corresponding operational specification that is defined and exploited for implementation purposes. The place of the two specifications with respect to the system can be seen in Figure 3.

Abstract Formal Specification of the Visual Interface

In this area, we started by considering the definition an abstract syntax and a formal semantics for the visual interface. The abstract syntax is intended to provide a snapshot of the visual environment in terms of static and compact structures. This visual syntax reduces distance to mathematical objects (such as some specific data mining functions). We intend to describe an abstract formal specification of the visual interface for each of the visual interaction environments corresponding to each of the data mining methods currently supported by VidaMine aka metaqueries, clustering, and association rules.

In order to define a semantics for a supported data mining method, it is necessary to first define the visual language supported by the corresponding visual interaction environment (user interface). An abstract syntax for a visual language can be defined in terms of multi-graphs as follows (Erwig, 1998). Let $\alpha, \beta$ be sets representing label types. A directed labeled multi-graph of type $(\alpha, \beta)$ is a quintuple $G(V,E, l, v, e)$ consisting of finite sets of nodes $V$ and edges $E$ where $l : E \rightarrow V \times V$ maps every edge to the pair of nodes it connects, $v : V \rightarrow \alpha$ maps every node to its label, and $e : E \rightarrow \beta$ maps every edge to its label. A visual language of type $(\alpha, \beta)$ is a set of directed labeled multi-graphs of type $(\alpha, \beta)$. Intuitively, visual controls are the nodes of the multi-graph, whereas the spatial relations between visual controls are the edges of the multi-graph. Since many of the visual controls appearing in the user interface are standard (such as check buttons, radio buttons, combo boxes, spin...
boxes, edit boxes, and sliders), their geometric properties will be ignored and simply use an abstract value to represent each visual control. Based on the foregoing notion of abstract syntax, we will define the visual language for each of the three visual interaction environments, and then describe the corresponding semantics.

Clustering

The clustering environment in VidaMine provides various visual widgets for specifying or selecting parameters characterizing a clustering task. The parameters include a fixed number of clusters or a measure (of homogeneity, separation, or density); attributes that will be directly involved in cluster analysis; and supplementary attributes (e.g., for labeling cases in the output). Specifying each such parameter may also involve more specific options/settings. The corresponding environment is seen in Figure 4.

In clustering, it is virtually impossible to apply uniformly a single technique to uncover the variety of structures present in multidimensional data sets (Jain et al., 1999). In most data mining applications, the user has to evaluate a set of techniques and then, based on experience and domain knowledge, select possibly more than one data mining technique to apply on the target dataset. A visual data mining system aimed at supporting the user during the entire mining process might be expected to provide support to the broadest range of clustering techniques possible. While it is true and commendable that VidaMine has an extensible framework, it is worth observing that adding one extension module for every clustering technique would render the system hard to utilize to anyone but the most informed practitioner, due to the variety of parameters and terminology. Therefore, in VidaMine, the approach is to provide and support three generalizations from the commonest types of clustering techniques. The three clustering generalizations are:

![Figure 4. The clustering environment: Input. Reproduced from Kimani et al. (2004) with the permission of Elsevier B. V.](image)

Table 1. Definitions of the main sets and functions

<table>
<thead>
<tr>
<th>Main Sets and Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset</strong> $S = {O_i</td>
</tr>
<tr>
<td>Symmetric dissimilarity function $diss : S \times S \rightarrow R^+$</td>
</tr>
<tr>
<td><strong>Classification</strong> $C$ of $S$ is a subset of a partition of $S$</td>
</tr>
<tr>
<td><strong>Accuracy function</strong> $m$ is a function on the set of all classifications of $S$ to $R^+$</td>
</tr>
</tbody>
</table>
1. Clustering based on homogeneity.
2. Clustering based on separation.
3. Clustering based on density.

Here are some common terms that will be used while describing the foregoing clustering generalizations: Consider a dataset $S = \{O_i \mid \text{for } i = 1, 2, \ldots, N\}$, and a symmetric dissimilarity function $diss : S \times S \rightarrow R^+$. It is worth noting that classification $C$ of $S$ refers to a subset of a partition of $S$. Let an accuracy function $m$ be a function on the set of all classifications of $S$ to $R^+$. The foregoing main sets and functions are summarized in Table 1.

The justification for adopting the three clustering generalizations (or categorization) is highlighted in the descriptions of the generalizations. In the sequel are the descriptions of the clustering generalizations.

**Clustering Based on Homogeneity or Separation**

Traditionally, clustering methods have been classified into a taxonomy having two most general groups: hierarchical and partitional (Jain et al., 1999). Hierarchical methods produce a sequence of nested partitions, whereas partitional methods only produce one partition. Such distinction, although relevant for many application domains, does not highlight similarities and differences between the various definitions of a cluster that are implicit in the methods.

Alternatively, clustering can be regarded as an optimization problem, in which the function to be optimized is a mathematical measure of homogeneity or separation (Hansen et al., 1997). Homogeneity is a global measure of the similarity between points belonging to the same cluster, separation is a global measure of the dissimilarity between points belonging to different clusters. For instance, the maximum cluster diameter over all clusters is a measure of homogeneity, whereas the minimum distance between objects in different clusters is a measure of separation.

In such a perspective, clustering methods are classified according to a taxonomy of homogeneity or separation functions, which expresses implicitly cluster definitions. Such categorization is most likely more effective in capturing different behaviors in practice. It therefore provides a more natural avenue for the process of selecting a clustering algorithm, which is most suited to a particular application or domain.

It is worth noting that the actual implementation of the search for an optimal solution is left entirely unspecified. Therefore, a suitable uniform abstraction can be given for a broad class of methods, ranging from classical hierarchical methods, to approximation algorithms, and a consistent parametric user interface between the clustering engine and the graphical user interface can be designed.

Therefore and following up from Section “The Proposed Clustering Taxonomy and Framework,” clustering can be formally defined as an optimization problem (Hansen et al., 1997) that takes one of the following forms:

**Problem $\Pi_1$:** Given an integer $K > 1$, find the partition $P$ of $S$ of size $K$ such that $m(P)$ is optimal.

**Problem $\Pi_2$:** Given a threshold $\theta \in R^+$, find the partition $P$ of $S$ of minimum (maximum) cardinality such that $m(P) \leq \theta$ ($m(P) \geq \theta$). In other words, sometimes $m(P)$ is optimal when large and at other times it is optimal when small.

In this research work, consideration is made only for functions $m$ ranging over the union of two families $H$ and $S$. The former family represents homogeneity functions and the latter represents separation functions. For a given data set $S$ and dissimilarity $diss$, define $H$ by:
**Equation 5**

\[ H_\omega(P) = \{ Q : O \in C(O) \ diss(O, O) | Q \in \{ \text{max}, \Sigma, \text{avg} \} \} \]

**Equation 6**

\[ H_c(P) = \{ Q : O \in C_h(O) \ diss(O, O) | h \in H_\omega(P), Q \in \{ \text{min}, \text{max}, \Sigma, \text{avg} \} \} \]

**Equation 7**

\[ H = \{ Q_c \in P \ h(C) | h \in H_c(P), Q \in \{ \text{min}, \Sigma, \text{avg} \} \} \]

and \( S \) by

**Equation 8**

\[ S_\omega(P) = \{ Q : O \in C(O) \ diss(O, O) | Q \in \{ \text{min}, \Sigma, \text{avg} \} \} \]

**Equation 9**

\[ S_c(P) = \{ Q : O \in C_s(O) \ s(O) | s \in S_\omega(P), Q \in \{ \text{min}, \text{max}, \Sigma, \text{avg} \} \} \]

**Equation 10**

\[ S = \{ Q_c \in P \ s(C) | s \in S_c(P), Q \in \{ \text{min}, \text{max}, \Sigma, \text{avg} \} \} \]

where \( C(O) \) is the cluster containing object \( O \).

Equation 5 defines a family of pointwise homogeneity functions, expressing that the homogeneity of object \( O \) can be defined as either the maximum (i.e., worst-case) dissimilarity to other objects in the same cluster, or the sum or average of all such dissimilarities. Likewise, Equation 6 defines a family of clusterwise homogeneity functions, expressing that the homogeneity of a cluster can be defined as the maximum, minimum, sum, or average of pointwise homogeneity of all its objects. Finally, Equation 7 defines a family of partitionwise homogeneity functions; the homogeneity of a partition can be defined as either the minimum (i.e., worst-case) clusterwise homogeneity over all its clusters, or the sum or average of all such homogeneities. Equation 8-Equation 10 provide analogous definitions for the separation function. Note, however, that the quantifier expressing worst-case pointwise or partitionwise separation is the minimum instead of the maximum, and that the quantifiers defining the separation of \( O \) extend to every object not in its cluster. Equation 5-Equation 10 induce a simple taxonomy with four levels into which functions \( m \) are classified. At the first level, homogeneity is separated from separation. Then, classes at lower levels in the taxonomy are separated according to the objectwise, clusterwise, or partitionwise quantifier.

**Clustering Based on Density Estimation**

In this section, \( diss \) is assumed to be a distance function and \( S \) to be a subset of a metric space \((X, diss)\). By elementary intuition, clusters can be regarded as regions of the object space where objects are located most frequently. Such simple analogy leads to approaches to clustering based on statistical techniques of non-parametric density estimation (Ankerst et al., 1999; Ester et al., 1996; Hinneburg et al., 1998; Schikuta, 1996; Silverman, 1986). The goal of density estimation is to fit to a data set \( S \) a density function of type \( X \rightarrow R^+ \). The implemented system supports clustering based on an important family of estimates, known as kernel estimators (Silverman, 1986). Functions in such family are defined modulo two parameters, the window width \( h \), and the kernel function \( \psi \). The value of the estimate at \( x \in X \) is obtained by summing, over all data objects, a quantity modeling the influence of the object. Influence is computed by transforming distance, scaled by a factor \( 1/h \), using \( \psi \):
Since the estimated density at $x$ should be large when the number of data objects which are located near $x$ is large, far data objects should have little influence. Thus $\psi$ is usually a probability distribution function, which decreases monotonically with distance from zero. Commonly used kernel functions are the Gaussian kernel, 
\[
\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right),
\]
and the Uniform kernel, 
\[
\frac{1}{2}(\text{sign}(x+1) - \text{sign}(x-1)).
\]
The parameter $h$ controls the amount of smoothing of the estimate, i.e. the sensitivity of the estimate to local structure in the data.

Once a density estimate is computed for a given kernel function, clusters can be defined in the following way. Let $A(O)$ be a neighborhood of $O$, and let $\alpha_{h,\psi}(O)$ be the data object (if it exists) nearest to $O$ in $A(O)$ having density greater than $\hat{\phi}_{h,\psi}(O)$:

\[
diss(\alpha_{h,\psi}(O), O) = \min_{O_j \in A(O)} \{diss(O_j, O) : \hat{\phi}_{h,\psi}(O_j) > \hat{\phi}_{h,\psi}(O)\}
\]
The tree roots in the forest induced by $\alpha$ are objects whose neighborhood does not contain objects having greater density. Therefore, the tree roots may be considered as representatives of the modes of the density estimate, and the descendants of a root are part of a single cluster. Formally, $\alpha$ induces a forest whose connected components are clusters: If $\sim$ is the equivalence relation generated by $\alpha$, then the clustering $P$ is the set of its equivalence classes:

\[
P = \{C : (\exists x_o \in S)C = \{x \in S : x_o \sim x\}\}
\]
The method can be further enhanced by introducing a notion of noise objects (Ester et al., 1996): Density at a noise object is less than a specified threshold parameter $\theta$. Noise objects are not part of any cluster, thus the method generates a classification instead of a partition:

\[
C = \{C : (\exists x_o \in S)C = \{x \in S : x_o \sim x \wedge \hat{\phi}_{h,\psi}(x) \geq \theta\}\}
\]

Abstract Syntax and Semantics of Clustering
The visual language is defined by the following:

\[
\alpha = \mathbb{R}^+ \cup \{\min, \max, \Sigma, \text{avg}\} \cup \{\pi_1, \pi_2\} \cup \{\text{Homogeneity, Separation, Density}\}
\]

\[
\beta = \phi
\]

\[
\]

Finally, let for brevity $hpq = v(\text{HomPartwQCom}), hcq = v(\text{HomCluswQCom})$,
User Interface Formalization in Visual Data Mining

\[ h o q = v(H o m o b j w Q C o m), \quad s p q = v(S e p P a r t w Q C o m), \quad s c q = v(S e p C l u s w Q C o m), \]
\[ s o q = v(S e p O b j w Q C o m). \]

The classification \( C \) of the dataset \( S \) is defined by:

\[ C = P_{h o m}(S) \text{ if } v(accurRad) = \text{Homogeneity}, \]
\[ P_{s e p}(S) \text{ if } v(accurRad) = \text{Separation}, \]
\[ C_{d e n}(S) \text{ if } v(accurRad) = \text{Density} \]

and the following hold

\[ |P_{h o m}(S)| = v(N C l u s S p i n) \]
\[ |P_{s e p}(S)| = v(N C l u s S p i n) \]
\[ |C_{d e n}(S)| = v(N C l u s S p i n) \]

\[ m_{h o m}(P) = \min_{c \subseteq P} \{ v(S m o o t h S l i d)_c \} \]
\[ m_{s e p}(P) = \min_{c \subseteq P} \{ v(S m o o t h S l i d)_c \} \]

\[ m_{d e n}(C) = \min_{c \subseteq C} \{ v(S m o o t h S l i d)_c \} \]

\[ m_{h o m}(P) \leq v(H o m S l i d) \]
\[ m_{s e p}(P) \geq v(S e p S l i d) \]
\[ m_{d e n}(C) \leq v(D e n S l i d) \]

\[ m_{h o m}(P) = \min_{c \subseteq P} \{ v(S m o o t h S l i d)_c \} \]
\[ m_{s e p}(P) = \min_{c \subseteq P} \{ v(S m o o t h S l i d)_c \} \]

\[ m_{d e n}(C) = \min_{c \subseteq C} \{ v(S m o o t h S l i d)_c \} \]

Metaqueries: In the metaquery environment of VidaMine, the user can specify patterns/relationships between or among data tables in an intuitive visual manner. The interface provides “hooks” and “chains” through which users can visually specify the relationships. By simply linking two attributes, the users indicate to the system that they are interested in metarules that have the two attributes related. Therefore, VidaMine enables users to visually construct metaqueries of interest. The left-hand side of Figure 2 shows the part of the metaquery environment that supports the foregoing specification.

Abstract Syntax and Semantics of Metaquerying

Here a simple abstract syntax for the metaquerying visual interaction environment is defined, together with its semantics, that is, the set of rules discovered by the system as instantiations of the metaqueries corresponding to the visual state of the interface when the “torch” icon is clicked (see Figure 2). For consistency with the visual interaction environments for the other mining techniques, the metaquerying visual environment
represents metaqueries using relation schemes instead of literal schemes, that is, using named attributes to denote table columns, instead of the usual logical variables.

The metaqueries, which are added to the pool when the “add pattern” button is clicked, are all constructible metaqueries, given the named and unnamed relations (i.e., named “X”) in the target space. Note that the links between handles which are represented in the target space do not have a counterpart in the “IF . . . THEN” expressions in the pool. In fact, such expressions only show the relation schemes, and not the literal schemes, that compose the head and body of a rule. Therefore, only the content of the target space defines multiple occurrences of variables in a rule. Since a user might click “add pattern” several times with different configurations of the target space, before clicking the “torch” button, for simplicity only one such configuration is considered in the semantics.

In the target space, rectangular frames enclosing attribute or relation names, and lines representing links connecting the frames (ignoring the handles) are nodes of the multi-graph. A node representing a frame is labeled by the relation name or attribute name it represents (i.e., the name appearing inside the frame). Attribute and relation frames may be adjacent, and attribute frames may be connected by lines. Therefore, for relations, attributes and links, two visual relations need to be represented: adjacent and intersecting. Adjacency will be assumed antisymmetric: A frame is adjacent to another if and only if the two frames share a horizontal edge, and the first frame is located above the second in the display.

In the following, let $U$ and $R$ be universes of attribute names and relation names, respectively, in the database. Let $V = \{X_1, \ldots, X_m\}$ be a countably infinite set of variable symbols, and $W = \{P_1, \ldots, P_r\}$ be a countably infinite set of predicate variable symbols. $V$ and $W$ are assumed disjoint and both disjoint from $R$. Let also $\zeta: V \rightarrow V, \lambda: V \rightarrow W$ be injective functions. In the sequel, $\zeta$ and $\lambda$ will be used to construct the literals of the rules from the frames.

The language is defined by the following:

**Equation 26**

$$a = R^* \cup U \cup R \cup \{"X"\}$$

**Equation 27**

$$\beta = \{\text{adjacent, intersecting}\}$$

**Equation 28**

$$V \supset \{\text{ConfSlid}, \text{SuppSlid}\}$$

The set $IRS$ of rules returned by metaquerying is defined as follows.

**Equation 29**

$$IRS = \{r \in RS : \text{conf}(r) \geq v(\text{ConfSlid}) \land \text{supp}(r) \geq v(\text{SuppSlid})\}$$

**Equation 30**

$$RS = \{r : (\exists \sigma)(\exists mq \in MQ) r = \sigma(mq)\}$$

**Equation 31**

$$MQ = \bigcup_{b \in L} \{b \leftarrow b \in L \cup \{b\}\}$$

**Equation 32**

$$L = \{P(X_1, \ldots, X_m) : (\exists n) P = \text{pred}(n) \land \text{isrel}(n) \land (\forall i \leq m)(\exists n') X_i = \zeta(n') \land \text{inschema}(n', n)\}$$

where

**Equation 33**

$$\text{isadj}(n,n') \iff (\exists e)l(e) = (n,n') \land e(e) = \text{adjacent}$$
Equation 34

\[ \text{intersects } (n,n') \Leftrightarrow (\exists e)(e) = (n,n') \land e(e) = \text{intersecting} \]

Equation 35

\[ \text{pred}(n) = \begin{cases} \lambda(n) \text{ if } v(n) = "X", \\ v(n) \text{ otherwise} \end{cases} \]

Equation 36

\[ \text{isconn} = \text{isadj}^e \]

Equation 37

\[ \text{isrel}(n) \Leftrightarrow (\exists n') \text{isadj}(n', n) \]

Equation 38

\[ \text{islink}(n) \Leftrightarrow (\exists n')(\exists n'')(\exists n_1)(\exists n_2) \text{isrel}(n_1) \land \text{isrel}(n_2) \land \text{isconn}(n', n_1) \land \text{isconn}(n'', n_2) \land \text{intersects}(n, n') \land \text{intersects}(n, n'') \]

Equation 39

\[ \text{inschema}(n, n') \Leftrightarrow (\text{islink}(n) \rightarrow (\exists n'') \text{isconn}(n'', n')) \land \text{intersects}(n, n'') \land (\neg \text{islink}(n) \rightarrow \text{isconn}(n'', n)) \]

and \( \text{isadj}^e \) is the equivalence relation generated by \( \text{isadj}^e \), that is, the smallest equivalence relation containing \( \text{isadj} \). Therefore, an equivalence class contains nodes corresponding to frames gathered in one relation scheme in the target space.

\( L \) is the set of literals defined by the visual configuration (Equation 32).

In each literal, \( P \) is the relation name enclosed in a frame, or a distinct predicate variable, if the name is “\( X \)” (Equation 35 and Equation 37). Every variable corresponds to an attribute frame that is connected to the relation frame, which names the literal, or corresponds to a link that intersects such an attribute frame (Equation 39). The set \( MQ \) of metaqueries is obtained from \( L \) by generating one metaquery for each literal, having the literal as head and the remaining literals as body. The rules \( RS \) instantiating \( MQ \) are defined by means of a class of substitutions. In Equation 30, \( \sigma \) is any substitution which, given a metaquery \( mq \), consistently replaces exactly every predicate variable occurring in \( mq \) with a relation name. Finally, the rule set \( IRS \) is defined as the set of all rules in \( RS \) that satisfy the support and confidence specified by the sliders \( \text{SuppSlid} \) and \( \text{ConfSlid} \).

Association Rules: Let \( \Omega \) be a set of items, and \( T \) be a database of transactions, that is, a list of elements of \( 2^\Omega \). Association rules (Agrawal et al., 1993) are implications of the form:

Equation 40

\[ I_{v_1}, \ldots, I_{v_m} \rightarrow I \]

where \( /I, I_{v_1}, \ldots, I_{v_m} \in \Omega \). If \( r \) is the foregoing association rule, and \( t = \{ I_{v_1}, \ldots, I_{v_m} \} \), then \( r \) has confidence \( \text{conf}(r) \) and support \( \text{supp}(r) \) in \( T \) defined by:

Equation 41

\[ \text{conf}(r) = \frac{|\{ t \in T : \{ I_{v_1}, \ldots, I_{v_m} \} \subseteq t \}|}{|T|} \]

Equation 42

\[ \text{supp}(r) = \frac{|\{ t \in T : \{ I_1 \cup I_2 \} \subseteq t \}|}{|T|} \]

In the association rule environment of Vida-Mine, there is the provision of “market baskets.” As seen in Figure 5, the association rule environment offers two baskets, the IF basket and the THEN basket. The IF basket represents items in the antecedent part of an association rule and the THEN basket represents items in the consequent part of the rule. Users may “drag and drop” items from the target dataset into the relevant baskets.
When through with the visual construction of a particular association rule, the users may empty both baskets into a pool of association rules that are of interest to them, and may go ahead and construct another rule.

**Abstract Syntax and Semantics of Association Rules**

The VidaMine system's support for association rules permits the specification of syntactic constraints (Agrawal et al., 1993), that is, listing items, which must appear in the antecedent or the consequent. Such items are dragged from the target space to the “IF” or “THEN” basket, and later combined to form the rule and put into a “pool” (of rules). The reader may refer to the association rule visual environment in Figure 5, and where the “pool” of rules is formed below the baskets.

It is assumed that the abstract graph contains one “item” node for every item occurring in the “IF” . . . “THEN” rule, labeled with the item’s name.

Assuming also nodes representing confidence and support (see Section “Abstract Syntax and Semantics of Metaquerying”), the set of association rules returned by the system is:

**Equation 43**

\[ IAR = \{ r \in AR : \text{conf}(r) \geq v(\text{ConfSlid}) \land \text{supp}(r) \geq v(\text{SuppSlid}) \} \]

**Equation 44**

\[ AR = \{ I_1, \ldots, I_m \to I : V_{THEN} = \{ n \} \to I = v(n) \land (\forall n \in V_{IF})(\exists i \leq m)I_i = v(n) \} \]

where \( V_{IF} \) and \( V_{THEN} \) are the sets of “item” nodes in the rule’s “IF” and “THEN” part, respectively.

**Operational Specification of the Visual Interface**

As part of implementation, our research effort also considers the provision of an operational specification. This specification entails defining a concrete syntax of mining tasks and an XML-based communication protocol. The concrete syntax of mining tasks is a high-level syntax of each of the tasks proposed in our task analysis and usability tests (Kimani et al., 2003). This syntax describes the click-streams that are allowed to occur during the user’s interaction with the interface. The syntax reduces distance to the specification of procedural objects. The XML-based protocol serves as a two-way link for

---

*Figure 5. Association rule environment: Input. Reproduced from Kimani et al. (2004) with the permission of Elsevier B. V.*
exchanging information between any applications that deal with mining aspects. It is interesting to note that the protocol idea can be exploited not only to import and incorporate new components into the system framework, but also to export the system’s components to external systems.

We start by describing the syntax of mining tasks. We then use the clustering subtask to demonstrate how the communication protocol has been defined.

**Concrete Syntax of Tasks**

Specifying a concrete, high-level syntax of mining tasks can serve at least three purposes. First, it highlights the most useful usage patterns, which guided the interface’s layout and design. Second, it permits an objective evaluation of the results of usability tests. Finally, it provides a design specification, which can be refined to yield the complete parsing structure of the interface at the level of information passed to and returned by the mining engines. The abstraction level will be such that terminal symbols represent interactions that can be regarded as elementary in the current

---

**Figure 6. Mining task**

- **MiningTask** → **ClusteringTask** | **MQTask** | **ARTask**
- **ClusteringTask** → **ClusteringInputTask** “PerformClustering” **ClusteringOutputObservingInteracting**
- **MQTask** → **MQInputTask** “PerformMQ” **RuleOutputObservingInteracting**
- **ARTask** → **ARInputTask** “PerformAR” **RuleOutputObservingInteracting**

---

**Figure 7. Syntax of clustering**

- **ClusteringInputTask** → **ClusteringTargetDataSubTask**
  - **ClusteringParametersSubTask** → **SpecifyConstraints** | **AssignAttributes**
    - “ForwardingTargetRelation”
    - “SpecifyConstraints” | **AssignAttributes**
- **ClusteringTargetDataSubTask** → **SpecifyParameter**
  - **SpecifyParameter**
- **SpecifyParameter** → **SelectAccuracyMeasure** | “SelectCaseLabel” | “SelectMiningAttribute”
- **SelectAccuracymMeasure** → **SelectHomogeneity** | **SelectSeparation** | **SelectDensity**
- **SelectHomogeneity** → “PushHomogeneityRadioButton” | **SetHomogeneity**
- **SelectSeparation** → “PushSeparationRadioButton” | **SetSeparation**
- **SelectDensity** → “PushDensityRadioButton” | **SetDensity**
- **SetHomogeneity** → **SetHomogQuantif** | **SelectHomFixedNum** | **SelectHomogThreshold**
- **SetSeparation** → **SetSepQuantif** | **SelectSepFixedNum** | **SelectSepThreshold**
- **SetDensity** → “SetKernel” | **SelectDenFixedNum** | **SelectDensThreshold**
- **SelectSepQuantif** SetSeparation | **SelectSepFixedNum** SetSeparation | **SelectSepThreshold** SetHomogeneity
state of technology (e.g., clicking buttons, setting a slider’s position, selecting an item in a combo box, setting the value of a spin box, etc.). The syntax will be defined using standard grammar notation, where terminal symbols are enclosed in double quotes. For brevity, only the first occurrence of every grammar rule, which is shared by two or more tasks has been written.

Figure 6 describes the mining task. Each subtask (clustering, metaquerying, association rules) comprises an input task, clicking the “torch” button to run the discovery algorithms, and an output task.

In the sequel we briefly describe the concrete syntax of clustering input and which can be seen in Figure 7. The task is divided into two consecutive parts, namely specifying target data and parameters, matching the upper and lower sections of the clustering interface. Specifying constraints means performing a sequence of manipulations of the controls in the top part of the specification phase, e.g., setting the number of attributes of a given type, choosing a relation, etc. (see Figure 4). Assigning attributes corresponds to dragging and dropping attributes from the resource relations to the “X” relation, which is subsequently dragged into the target space. Selecting an accuracy measure leads to selecting either homogeneity, separation, or density, by pushing the respective radio buttons (see Figure 4). Subsequently (see Figure 7), the user may select quantifiers, or the kernel function and the amount of smoothing. Selecting the problem type gives access to setting the number of clusters or the threshold for the accuracy function.

An XML-Based Communication Protocol

The large amount of recently proposed mining techniques, and the need to easily select the most effective method for the multivariate data set at hand, calls for the design of extensible data mining systems. Exchanging XML documents allows for the addition of new mining engines or new methods for existing engines as plug-ins. Such a protocol would serve as a two-way link for exchanging information between any applications that deal with mining aspects. It is interesting to realize that this idea can be exploited not only to import and incorporate new components into the system framework, but also to export the system’s components to external systems. In the VidaMine system, communication between the GUI and the mining engines is established by means of XML document streams. The communication protocol is specified by XSchema for the selection of methods and passing parameters, and for passing output results to the GUI. Since PMML, the predictive model markup language (DMG), has already set an XML standard for representing mining results in the data mining community, our specification is based on, and in some parts extends, PMML.

Example of XML-Based Protocol for Clustering Task

Assuming the target dataset in Figure 4, the generated XML document for method selection is listed in Figure 8. Each attribute listed in the “X” relation matches exactly one corresponding “DataField” in “DataDictionary” and one “MiningField” in “MiningSchema.” In this research work, “active” corresponds to the “checked” state whereas “supplementary” corresponds to the “unchecked” state of the attribute’s check box in the target space. It is worth observing that “checked” attributes are those that have been chosen to directly participate in cluster analysis. “Unchecked” attributes are those that will not directly participate in cluster analysis but will be included in the results for other purposes (such as for labeling). “AccuracyFunction” allows for selecting one of “Homogeneity,” “Separation,” and
“Density.” Within each, the problem’s form can be selected by “FixedNumberOfClusters” (problem $\Pi_1$ of Section “Clustering Based on Homogeneity or Separation”) or “FixedThreshold” (problem $\Pi_2$ of Section “Clustering Based on Homogeneity or Separation”). The “Density” element includes the type of density estimate used. Currently, the only allowed type is “Kernel.” Other types can be easily added as alternatives to “Kernel.” Figure 10 is a view of the XSchema, which specifies documents containing the selection of a method and parameters for clustering. It is worth noting that the XSchema in Figure 10 does not allow for specifying the algorithm used to solve the problem.

Thus, for given problem definition and form of the accuracy function, the clustering engine may implement an exact or approximate algorithm. In the current example, the selection of min for objectwise, clusterwise, and partitionwise quantifiers is equivalent to the worst-case split clustering problem (Hansen et al., 1997). For such problem, an exact solution can be obtained applying the well-known single link hierarchical algorithm to the target dataset, halting at the topmost partition whose worst-case split value is less than the value of the split function (0.03 in the example). The
document containing clustering result/output, whose view can be seen in Figure 9, represents the solution to the clustering problem in Figure 4. For brevity, the XSchema for the clustering output is omitted, which is an extension of the corresponding XSchema in PMML. The “Cluster” elements describe five clusters named 1, 2, 3, 4, and 5. Each “Cluster” element contains the clusterwise accuracy function that was selected for clustering, and its value at the cluster. (The split of the cluster, in the example.) The engine may include any other accuracy function in the result document, specifying its value at the cluster, and the clusterwise parameters defining it. (In the example, the engine returned the cluster’s “Homogeneity” computed as its diameter.) Finally, “ClusteringArray” assigns each object in the target dataset to its cluster.

Mapping Between Clustering Task and XML Documents

It is important to define a mapping between the clustering and the XML-based communication protocol that was just defined. When the formal specification of the interface is defined through an abstract syntax and semantics, a mapping between user actions and XML documents describing the actual parameters passed to the engines can be defined by means of a recursive function. In the VidaMine system, a mapping function \( M \) is defined by recursion on the grammar rules describing the clustering task of Figure 7 to XML documents conforming to the XSchema of Figure 10.

Although for brevity we have omitted its formalization, it should however be noted that \( M \) basically is a mapping satisfying a list of equalities of the forms:

**Equation 45**

\[
(term)^M = xml(value_1, \ldots, value_p)
\]

**Equation 46**

\[
(nt_1, nt_2)^M = (EXT(nt_1)^M(nt_2)^M)
\]

In **Equation 45**, \( term \) is a terminal symbol, \( xml \) is a function constructing an appropriate XML element corresponding to \( term \), \( value \) is the value of a visual control. In **Equation 46**, \( nt \) is a nonterminal symbol representing the manipulation of a visual control, \( EXT \) is a function that extends the XML tree in its second parameter using the elements in the first parameter. \( EXT \) also provides appropriate defaults when the user’s actions do not affect some of the controls.
Comparison with Data Mining Systems

SQL Server 2005, developed by Microsoft Corporation, includes several data mining tools: association rules, clustering, decision trees, linear and logistic regression, naive Bayes, and neural networks. Its interface is based on menus, panes, and tabs. Three panes present the user with an explorer tree of folders related to analytical tasks (data sources, multidimensional cubes, mining structures, etc.), a properties visualization tool, and tabs for model settings, visualization, evaluation, and for prediction. Parameters for the mining algorithms can be accessed from pop-up menus. The visualization tab supports all mining models with several types of visualization each, with some active graphical elements. Visual features, such as color, can be associated to quantitative features of models, such as relative frequency. Overall, the interface provides a powerful set of functionalities, but requires training to be used, in particular for structuring the mining task. No formal model is known for the graphical part of the interface.

Clementine was originally developed by Integral Solutions Ltd (ISL) and later distributed as a part of the SPSS system, after SPSS purchased ISL on December 31, 1998. Clementine supports many prominent mining techniques including clustering, association rules, sequential patterns, factor analysis, and neural networks. Users construct a map (a “stream”) of their data mining project/model by selecting icons called “nodes” that represent steps in the data mining process. Clementine allows users to adjust/refine their “streams” and rerun the system on the refined model. Although its visual interface illustrates the flow of control and data, allowing the user to better understand and follow the mining process, no formal syntax and semantics of the interface as a whole is reported.

Enterprise Miner, developed by the SAS Institute, provides a large number of data mining algorithms, including decision trees, neural networks, regression, radial basis functions, self-organizing maps, and clustering. Enterprise Miner has an interesting visual interface, allowing for powerful support of data transformation, visualization tools for multidimensional analysis, and extensive parameter options for the algorithms. The user drags and drops icons from a task bar into a work area where the icons can be connected. Each connection represents the flow of information from one node to its successor, and each icon type represents a type of processing (for example, imputation of missing data, declaration of variable types and target variables, regression, etc.) on the data from the predecessor node. Mining algorithms are viewed as a special type of processing. The visual metaphor of nodes and edges is intuitive; however, most of the complex settings of each node type cannot be operated upon graphically, but must be accessed through a type-specific tab-based interface. The effort does not report any formal description of the graphical interface available.

Intelligent Miner is a data mining software developed by IBM, supporting association rules, clustering, classification trees, and neural networks. There are two versions of Intelligent Miner, Intelligent Miner for Data and Intelligent Miner Modeling/Scoring/Visualization. Intelligent Miner for Data is a standalone program. Its interface is based on explorer-like panes and drop-down menus introducing the various data mining techniques, parameter settings, and visualization windows. Intelligent Miner Modeling/Scoring/Visualization is a suite of three distinct tools, for the extraction mining models, for prediction, and the visualization of models. The product is integrated into the IBM DB2 Universal Database, which provides all metadata support, for example by storing the mining models and their visualizations. The interface of the modeling component is textual and consists of stored procedures for data mining algorithms callable from DB2’s command windows, whereas the visualization component
has a graphical display with active elements and extensive sorting options. The interface of Intelligent Miner for Data has no known formal specification, as the interface of Intelligent Miner Modeling/Scoring/Visualization. (For the latter, a formal specification is relevant only to the visualization component, which is however, the same supplied with the standalone version.)

OPEN ISSUES AND FUTURE TRENDS

Despite the remarkable efforts and developments that have taken place in the field of visual data mining, the field still faces quite a number of challenges and also exhibits various issues as highlighted in the sequel:

• Data mining methods normally can handle or cope with large amounts of data. Since visual data mining systems integrate data mining techniques and visualization techniques, visualization techniques have to face up to the challenge regarding large quantities of data. Bearing in mind the limitations of the display space, visualizations have to seek effective mappings so as to display all relevant data. Moreover, visualizations have to use efficient techniques to load relevant data in the main memory.

• Most of the existing visual data mining systems have not been subjected to thorough usability evaluations. In order to determine the usefulness of visual data mining systems, the importance of usability studies cannot be overemphasized. On the same note, it is worth acknowledging that the lack of precise theory or a universal/standard framework for developing visual data mining systems may not only be partly responsible for the current situation, whereby each visual data mining system seems to be an ad-hoc development, but may also to some extent at least complicate the usability issue as far as the visual data mining field is concerned. In that respect, Grinstein, Hoffman, Laskowski, and Pickett’s work on benchmarking and evaluation in visualizations for data mining (Grinstein et al., 2002) might be instructive.

• It is a challenging activity to find the methods, techniques, and corresponding tools that are suitable for a specific visual mining task, or a particular type of data. One way forward would be to evaluate the disparate frameworks to determine their effectiveness and to verify their applicability in different application domains (Simoff, 2001).

• Visual support often necessitates the usage of graphical elements. The user is expected to interact with these graphics in real time. When the user interacts with the visualization, there should be some kind of immediate feedback from the graphic. Bearing in mind the fact that the memory demands on the computer that are posed by graphics, may decrease the performance of the system. Consequently, changes introduced on the visual environment may not immediately reflect the changes occurring in the information realm in real time and vice versa. It is no trivial task determining a good balance between interactivity and performance.

In terms of our future research activities, we are planning to support the output of descriptive statistics, principally using pie charts and histograms, in the clustering environment. It also is worth noting that a set of the system features are under an engineering process in order to become part of the commercial suite D2I, distributed by Inspiring Software (http://www.cisat-group.com). The D2I (Data to Information) environment will serve as a generic suite for handling the data to be subjected to mining. Even as we consider the foregoing research directions, it is also important to be aware of various trends in data mining. One
of the current trends is distributed data mining, where the target data is located in different places and/or in different physical locations. There is also a growing interest in ubiquitous data mining, where the user may carry out data mining from ubiquitous devices (such as PDAs, cellular phones, etc) rather than from the conventional computer. Another area that has been drawing a lot of attention is hypertext and hypermedia data mining. Efforts in the foregoing type of data mining would normally entail one or more of the following three aspects: content mining, structure mining, and usage mining. It is also interesting to observe a growing interest in audio data mining, where audio signals are used to represent data trends/patterns or other specific data features.

CONCLUSION

Despite the existence of many data mining efforts that deal with user interface aspects, there is relatively little work going on or that has been published on the specification of the syntax of such user interface and the corresponding semantics. In general, a formal specification can bring about many benefits such as facilitating the description of the system properties without having to be concerned about implementation details and enabling the detection of fundamental design issues before they manifest themselves in the implementation. In visual data mining, where visualization is often a key ingredient, a formal specification can be rewarding. For instance, it can enable users to decide which interaction/operation to apply to get a particular output, it can help users to predict the output of their interactions/operations with the system, and it can facilitate the development of interaction models. In this work, we have proposed and described an approach for specifying such a formal specification in the process of developing a visual data mining system, VidaMine.

REFERENCES


Kimani, S. (2002). An effective visual data mining environment. *Doctoral Posters at the International Conference on Very Large Data Bases (VLDB)*.


Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Proceedings of the International Conference on Very Large Data Bases (VLDB)* (pp. 144-155). San Francisco: Morgan Kaufmann Publishers.


User Interface Formalization in Visual Data Mining


ENDNOTE

1 VidaMine is an acronym for VIusal DAta MIINing Environment.
Chapter 2.30
Understanding the Nature of Task Analysis in Web Design

Rod Farmer
The University of Melbourne, Australia

Paul Gruba
The University of Melbourne, Australia

ABSTRACT
Designing usable Web-based interfaces challenges practitioners to carefully consider end-user behaviour and requirements. Unfortunately, in meeting this challenge, human-computer interaction (HCI), task analysis is often poorly understood and applied during Web design activities. Rather than purely evaluating usability against prescriptive guidelines, we argue that designing for Web-based interaction requires a more holistic and descriptive approach. This chapter provides an overview of cognitive and postcognitive HCI task analysis frameworks, and their respective abilities to capture a systemic view of stakeholder requirements. As such, this chapter provides a valuable resource for researchers and practitioners alike.

INTRODUCTION
Although improved system design results when researchers and developers understand how users use technology (Raeithel & Velichkovsky, 1996), understanding individual user traits, such as motivation and other contextual factors that guide user participation during computer-mediated activities can be deceptively complex. Simply asking users what they want and how they use a system is further complicated by the fact that users are often incapable of vividly and objectively describing their experiences with the system (Sommerville, 2004). Expertise, sociocultural, and organisational policy factors may impact perception of purpose, meaning, and context, and hence influence the quality of user feedback (Gasson, 1999). Therefore, determining whether a
Understanding the Nature of Task Analysis in Web Design

system is fit-for-purpose for a particular end-user population can be extremely challenging.

As developing fit-for-purpose systems is a principal concern of human-computer interaction (HCI), the system design process must ensure that end-user requirements are validated against those who have a vested interest in its use (stakeholders). Therefore, choosing the right HCI framework for eliciting, analysing, and modelling stakeholder requirements is critical for ensuring overall system quality (Farmer, Gruba, & Hughes, 2004). The process of seeking to understand the human nature of these requirements is referred to in HCI as task analysis. There is a wealth of frameworks, models, methodologies, and tools that can be applied to assist in this process. However, choosing the “most appropriate” approach is dependent upon several factors, including: the domain, context of use, and available resources.

Task analysis is arguably the most important aspect of HCI as it provides the analyst, researcher, or developer with insights into the nature of human behaviour. A major benefit of conducting task analysis throughout the software development life cycle (SDLC) is its communicative power and ability to elicit and elucidate requirements throughout each phase of development via a set of formalised attributes and notations. Unfortunately, comparing and choosing the right task analysis approach during system design is frequently hampered by the lack of universal notations and user attributes that can be applied across frameworks (Balbo, Ozkan, & Paris, 2004).

The primary aim of this chapter is to provide a critical overview of task analysis in HCI and its application to Web design. Specifically, the chapter will discuss the cognitivist origins of task analysis, and the recent shift towards more ecologically valid approaches. We discuss several leading approaches within each paradigm and describe their general applicability to Web design. We conclude with an integrative approach to task analysis that attempts to bridge the divide between cognitive and postcognitivist perspectives.

**TASK ANALYSIS IN HCI**

The term task analysis is commonly used to denote a wide range of activities and processes that attempt to either describe, equate, or predict human performance during task-based interaction (Diaper, 2004). A direct corollary of early cognitive psychological research concerning cognition and procedural knowledge (Kirwan & Ainsworth, 1992; Miller, 1953, 1962), task analysis has been applied successfully to numerous fields of research, including:

- Interactive system design (Newman & Lamming, 1998)
- Safety critical systems design and evaluation (Paternò & Santoro, 2002)
- Cognitive engineering (Rasmussen, Pejtersen, & Goodstein, 1994; Vicente, 1999)
- Computer-assisted language learning (Corbel, Gruba, & Enright, 2002; Farmer & Hughes, 2005a, 2005c)
- Multi-modal interaction (Farmer, 2005)
- Intelligent learning object classification (Farmer & Hughes, 2005b)
- Social intimacy (Vetere et al., 2005)
- Web design (Dix, 2005)

It is therefore not surprising to see an increasingly divergent array of theoretical perspectives emerging on the nature of human-machine interaction. Observing that methodologies in HCI have already reached a sufficient level of sophistication and application, Whittaker, Terveen, and Nardi (2000) have argued that it is time to address existing problems, rather than develop additional idiosyncratic models and notations. Indeed, renewed focus has recently been applied to the problem of integrating and grouping task analysis theories and techniques, which at first glance may appear fundamentally incommensurate (Farmer, 2006; Wild, Johnson, & Johnson, 2003).

The primary aim of task analysis is to produce a reliable procedural description of human praxis.
Understanding the Nature of Task Analysis in Web Design

Table 1. Types of task analysis (MITRE, 2003)

<table>
<thead>
<tr>
<th>Types of Task Analysis</th>
<th>System evaluation methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive task analysis</td>
<td>Descriptive analysis</td>
</tr>
<tr>
<td>Knowledge elicitation</td>
<td>Human reliability analysis</td>
</tr>
<tr>
<td>Computational cognitive modelling</td>
<td>Cognitive-oriented methods</td>
</tr>
<tr>
<td>Task analysis</td>
<td>System-oriented methods</td>
</tr>
<tr>
<td>Computational task simulation</td>
<td></td>
</tr>
</tbody>
</table>

However, unlike mere procedural functions, tasks incorporate the notion of purpose (Preece, Rogers, Sharp, Benyon, Holland, & Carry, 1994, p. 411), hence goals and planning. Many task analysis techniques are built on the belief that systems are reducible to their constituent parts, promoting a systematic, linear decomposition of human praxis (Watts & Monk, 1998). However, systems can never be logically decomposed into subsystems without losing some implicit value, whether operational (knowledge) or functional (capability) (Latour, 1987). This said, linear reductionist descriptions of work may be appropriate for developing highly constrained, well-defined environments, such as interactive voice response (IVR) systems that typically restrict users to inhabiting a single state. Such cognitive environments therefore promote a strong adherence to a narrow view of user behaviour (Robinson, 1993). Table 1 shows a variety of task analysis approaches typically employed within HCI.

Disregarding the actual methods applied within these frameworks, we can further collapse these approaches onto two conceptual axes: descriptive/analytical and cognitive/system. The descriptive/analytical axis reflects whether the purpose of the analytical framework is to describe or to empirically analyse interaction behaviour. The cognitive/system axis reflects whether the framework is cognitively oriented, concerned with modelling mental processes during work activity, or system oriented, concerned with modelling how individuals relate cognitively to their social, organisational, and environmental work contexts. As these axes represent continuums, more fine-grained categories of task analysis now emerge.

Low-level analytic/cognitive approaches such as goals, operators, methods and selection rules (GOMS) and hierarchical task analysis (HTA) tend to treat task interaction as a series of linearly sequenced actions from which we can derive the sense of the activity, and thus effective system design. Higher-level system/descriptive approaches, such as soft systems methodology (Checkland, 1999), tend to take a more holistic perspective, focusing on situated activity rather than actions. The growing importance of modelling situation awareness during HCI is representative of a paradigmatic shift away from individual, information processing accounts of interaction, towards more socially relevant, tool-mediated representations of activity.

Gathering Requirements via Task Analysis

Requirements engineering (elicitation, analysis, and specification) is essential to developing quality Web-based systems as it (1) helps to elicit possible user groups and their level of involve-
Understanding the Nature of Task Analysis in Web Design

ment; (2) provides insight into user and system requirements; and (3) places the development effort within a contextual environment (Balbo et al., 2004; Farmer et al., 2004). Schach (1999) provides a case-based analysis of software development projects, highlighting that the origin of 80% of all software defects can be traced back to earlier requirements engineering and design phases. In addition, the cost of fixing a defect during the maintenance phase, as opposed to these earlier phases, is an order of magnitude. As both requirements engineering and task analysis serve to determine which features of a system will render it fit-for-purpose, it is necessary to ensure they are conducted appropriately throughout the SLDC (see Table 2).

According to Mager (1991), there are four distinct fields of work analysis: (1) performance, (2) critical incident, (3) task, and (4) goal analysis. Performance analysis attempts to describe aspects of human activity that are directly related to understanding and improving human performance. Traditionally, this area of research has focused on two aspects of work: operator awareness and capability modelling; and human-system simulation environments. Operator modelling is for the most part a behavioural activity, using techniques such as time analysis and operator function models (Kirwan & Ainsworth, 1992). Human-system simulation environments, on the other hand, require sophisticated, computational-task simulation tools and theories to evaluate human decision making, including recognition-primed decisions and naturalistic decision making (Zsambok & Klein, 1997).

Critical incident analysis focuses on developing causal models of relationships within complex systems to prevent accidents or error states in safety-critical environments (Shrayne, Westerman, Crawshaw, Hockey, & Sauer, 1998).

<table>
<thead>
<tr>
<th>Phase</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery and definition</td>
<td>To elicit, analyse, and specify functional and nonfunctional stakeholder requirements within the context of use and the existing limitations and constraints upon the activity To generate high-level, coarse-grained descriptions of the main tasks and objectives that are relevant to the user(s)</td>
</tr>
<tr>
<td>Design</td>
<td>Traditionally, the role of this phase has been to develop the user-interface design, however it should also include high-level, system architecture specification, documentation and additional resource design. These activities should all occur concurrently. To define and model the generic functionalities of the system, especially consistency and learnability (affordance, usability) when the system is to be deployed across several platforms</td>
</tr>
<tr>
<td>Development and Deploy-</td>
<td>To analyse the implemented functionality of the system in terms of efficiency To automatically generate parts of the architecture and subsystem components related to the functional stakeholder requirements</td>
</tr>
<tr>
<td>Evaluation and predi-</td>
<td>Through the use of specific task modelling notations, to produce a series of design tests which evaluate user performance and ensure that the final product is fit-for-purpose</td>
</tr>
</tbody>
</table>

Table 2. Purpose of task analysis various stages of the SDLC
As such, this work requires a higher-degree of probabilistic modelling and stochastic analysis than the other fields. More than any other work analysis related field, critical incident analysis attempts to reduce complex activities to discrete, measurable events.

Perhaps the most difficult category to define, however, is task analysis. Researchers and practitioners have failed to agree upon a universal definition of “task,” including an appropriate unit of analysis (Carroll, 2002). As a result, the term task analysis has been used to describe all four categories of work analysis, especially goal analysis. Accordingly, contention has arisen concerning the structure of tasks, and their relationship to goals and planning (Nardi, 1996; Suchman, 1987).

What is a Task?

Figure 1 depicts a traditional structural view of task in HCI where it is seen as a conscious act (Kirwan & Ainsworth, 1992; Watts & Monk, 1998). Task-as-activity is comprised of some goal that must be achieved through mediated interaction via agents and artefacts of the environment (Flor & Hutchins, 1991; Hutchins, 1995). Planning is implied as goals reflect the system state to be achieved through effecting some combination of actions (events). Unfortunately, here we run into trouble. Preece et al. (1994) conceive activities as devoid of control structures and not requiring thought. While this may be true of some activities where procedural knowledge has reached a certain level of automaticity through habituation, it cannot be true of all interactive systems, such as Web environments.

Goals require constant monitoring and evaluation (Suchman, 1987), which implies conscious decision making. Indeed, it is more logical to assume that tasks are constructed and maintained within a frame of reference that includes planning (Filkes, 1982). Planning requires knowledge of when and how to initiate tasks. We therefore question the correctness of this traditional view of Task, and propose Figure 2 as an alternative. Here the term task is substituted for activity. As activities are motivated social constructs (Leont’ev, 1978), it is appropriate that they appear superordinate to goals, individual goals being a subset of collective goals. As activities exist at the social level, they are less temporal than subjective individual acts. Planning is therefore rooted firmly within specific communities of practice, and less open to change and violent fluctuations in behaviour. Given their degree of permanence, we classify conscious decision making conducted at this level, meta-planning. Furthermore, Suchman (1987) argues that individuals plan reflexively rather than deliberatively, responding to immediate changes in environmental conditions. As such, planning at the goal level is likely to be affected by micro changes during an activity. We classify this within-activity planning, micro-planning. Figure 2 suggests that tasks are subordinate to goals, thus inverting the traditional view. Moreover, Figure 2 distinguishes between two types of task: conscious and unconscious.

The mediating variable in task classification is habituation. The more one practices an event, the fewer attentional (cognitive) resources need to be assigned to performing that task. Conversely, when an error or breakdown in information flow occurs, these tasks are raised into consciousness.

Figure 1. Traditional structural view of task

| Task       | Conscious
|------------|--------|
| Goal       | Planning
| Actions    | Events |
Understanding the Nature of Task Analysis in Web Design

Divisions within HCI on Task Analysis

The notion of Task has historically been dependent upon the predominant theoretical models of the time. According to Goguen (1996), there are today two dominant theoretical perspectives within HCI: the cognitive (cognitive science and experimental psychology) and the postcognitive (sociology and anthropology). These perspectives exist along a continuum polarised by the split between the natural and social sciences (Raeithel & Velichkovsky, 1996). Selecting between cognitive and ecological theories of human praxis can have considerable impact upon system development. For instance, the view that participation within an activity is not isolated, yet inexorably bound to external relationships and sociocultural conditions poses serious questions for how we model both functional and nonfunctional requirements during online system design.

Cognitive approaches to task analysis primarily focus on aspects of user performance (Diaper, 2004), often neglecting sociocultural phenomenon essential to meaningful interaction. Postcognitivist, or ecological, approaches underscore the importance of context in evaluating human praxis. Postcognitive approaches hold that meaningful interaction can only be derived from careful consideration of the environmental, sociocultural, and historical contexts in which an activity occurs. Moreover, activity itself is distributed among various cognitive systems within the environment, and not just at the user interface, as interaction requires the accessing of information and/or anticipation of events that may be beyond the user’s experience. Hence, measuring the meaningfulness of interaction is
not merely an empirical measure of performance (human or technological), rather an evaluation of the appropriateness of an activity in relation to a set of established goals. Furthermore, meaning during online activity is negotiated through the user interface, and not at the user interface.

Contrary to the more behaviouristic views of cognitivist approaches, postcognitive perspectives are driven primarily by sociological considerations, including how work is shared, organised, and completed within communities of practice (Wegner, 2003). This shift is also indicative of increased research into how we can better capture the context of interaction; develop greater understanding of user behaviour — in situ —; and how tool mediation impacts knowledge acquisition and management (Kaptelinin, 1996b).

The cognitivist dialogue maintains an assumption that derived task models can be universally applied across different user groups, implying a greater degree of similarity between individuals than should otherwise be accorded (Preece et al., 1994). Carroll (1997) suggests that designing systems upon the notion of universally applicable task models implies certain cognitive equivalences between humans and computers, yet maintains that artefacts lack self-awareness, and should never be thought to possess knowledge of their behaviour. Although some may suggest that artefacts possess certain cognitive affordances, this is often a misinterpretation of the impact of social process, such as cultural conventions, on object use (Norman, 1999).

APPROACHES TO TASK ANALYSIS IN HCI

As previously highlighted, the two primary task analysis movements within the field of HCI are cognitive (behaviourist) and postcognitive (socio-cultural). Methods employed within the cognitivist paradigm promote action-oriented analyses of interaction (Kaptelinin, Nardi, & Macaulay, 1999). Postcognitive frameworks represent not so much an opposing view, but rather an argument for greater emphasis on modelling human activity as it occurs in situ. Postcognitivist methodologies favour goal-oriented or object-oriented analyses. To better understand how these paradigms differ from one another, we now consider related methodologies.

Behavioural Task Analysis

Behavioural, or cognitive task analysis frameworks focus primarily on how human behaviour is determined by the user’s internalisation and transformation of input, and how this information results in output as action. Cognitive approaches to HCI mostly focus on the user-computer interface, as it forms the critical point of coherence between two discrete models of processing — that of the human and the computer. While this is true of most underlying cognitive task models, differing purposes of use have resulted in considerable variation between approaches.

Cognitive Task Models

There are various levels of behavioural task analysis approaches within HCI. These include low-level cognitive methodologies, such as HTA (Annett, 2004; Annett & Duncan, 1967; Shepherd, 2001), and higher-level, meta-cognitive frameworks, such as scenario-based design (Carroll, 1996, 1997, 2000, 2002; Rosson & Carroll, 2002).

Behavioural frameworks are structured upon specific task models, structured descriptions of the interactive task to be performed by the user(s). It encompasses the notion of task and as such represents the specific set of actions that the user may undertake in a goal-directed activity. Task models may also include a task hierarchy, indicating task sequencing and any additional constraints upon usage (Limbourg & Vanderdonckt, 2004). Although most vary substantially in their depth
of analysis, degree of formalism, and purpose, all task models comprise a set of common objectives (Table 3).

According to Table 3, these common objectives imply that task models can be used for a variety of purposes, from requirements engineering through to system evaluation, representing varying levels of abstraction and serving different purposes within the SDLC (Balbo et al., 2004).

The underlying theme within behaviourist frameworks is that users have an overarching system goal they wish to achieve, which through intentionality is transformed into action. Decision making occurs through comparison of intention and perceived system behaviour (feedback) (Norman, 1999). As such, interaction is modelled as a closed information-processing loop — input, transformation, and output. These types of analysis are effective for modelling and evaluating the primitive tasks that determine user behaviour in well-structured activities. Here, we briefly describe a set of highly influential, cognitive task analysis frameworks. With the possible exception of cognitive work analysis (CWA), we argue that these approaches fail to sufficiently support decision making and problem solving in unstructured and dynamic environments.

### Goals, Operators, Methods and Selection Rules

GOMS-based approaches have received considerable attention in the HCI literature (John & Kieras, 1996a, 1996b), as it can be used quantitatively to predict efficient user interaction with a system design (task performance), and qualitatively as a means of modelling low-level tasks (Kieras, 2004). GOMS-based approaches are also effective at predicting future task performance. However, models of interaction derived from a GOMS analysis tend to be abstract, rather than specific. Moreover, GOMS — as initially conceived — does not account for motivational factors involved in computer-mediated activities, nor is it well suited for modelling social and collaborative work practices.

The unit of analysis in GOMS is the task. Like HTA, GOMS describes task-based interaction as a hierarchically ordered sequence of actions, aligned with specific goals (and subgoals) (Card, Moran, & Newell, 1983). Briefly, GOMS views tasks as comprising:

1. **Goals**: Desired outcomes within some computer-mediated activity;

### Table 3. Common objectives of task models (Bomsdorf & Szwillus, 1998)

<table>
<thead>
<tr>
<th>Objective</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>To inform</td>
<td>To inform users and designers about potential problems. Common themes involve needs analysis, usability, and affordance.</td>
</tr>
<tr>
<td>Evaluation of human perfor-mance</td>
<td>To establish the impact of new or existing tools upon task performance in some work practice</td>
</tr>
<tr>
<td>System design support</td>
<td>To provide a detailed and structured conceptual model of the task, including behavioural (time, errors, feedback, etc.) and structural modelling (functionality, visibility, etc.)</td>
</tr>
<tr>
<td>System generation</td>
<td>To help develop prototypes, user interfaces, and elements of the system architecture</td>
</tr>
</tbody>
</table>
Understanding the Nature of Task Analysis in Web Design

2. **Operators**: Actions that must be executed in order to achieve some goal;

3. **Methods**: The set of operators that are required to achieve some larger (system) goal; and

4. **Selection rules**: Logic rules used to choose between competing methods given some system feedback.

Within GOMS, input is received from the system, which is processed and associated with internalised representations of the system (mental models). This internal system information is aligned with immediate and overarching goals, and subsequently associated with competing possible actions. Selection rules serve as transformation functions, helping to determine which actions possess the greatest utility, and thus provide the most effective means of interaction. Output is the corollary of some cognitive utility evaluation, representing the most effective path to achieving the desired goal.

For GOMS, generating meaning from interaction occurs at the man-machine interface and is highly determined by subjective internal representations of the activity. Unfortunately, GOMS is highly prescriptive and fails to consider distributed models of interaction, which consider environmental artefacts, organisational behaviour, and social conventions that reflect historical decision-making processes and cultural conventions (Hutchins, 1995).

**Hierarchical Task Analysis**

HTA is an action-oriented, decompositional framework primarily concerned with the training of people in system design and use. Initially developed for analysing nonrepetitive operator tasks (Annett, 2004; Annett & Duncan, 1967), HTA grew out of the need for a more expressive model for describing mental processes during man-machine interaction. HTA's theoretical foundations lie in system theory and information processing, seeing task performance “as the interaction between human and machine, the latter becoming increasingly complex as computers and automation developed” (Annett, 2004, p. 68). With relation to task performance, the term *analysis* in HTA refers to the process of problem identification and structuring. HTA is an effective process for proposing empirical solutions to existing specified problems (Annett, 2004).

In HTA, *goal* refers to an expected outcome, or system state. Goals and subgoals, may be active or latent, arising when the need presents itself to achieve some expected outcome. HTA frequently models this goal-oriented behaviour as tree structures, such as decision trees. Goals are established and acquired through an information processing cycle similar to that in our discussion of GOMS (Annett & Duncan, 1967; Kieras, 2004). As with goals, *actions* in HTA may also be decomposed into nested actions, each maintaining their direct relationship to an established system state. Being nested, actions are available at both the current node of activity, and at their super node (parent task). Therefore, according to HTA, user behaviour can be reduced to a cycle of monitoring for new input, deciding upon available alternatives and controlling subsequent behaviour.

Although an attractive framework, there are a number of problems HTA practitioners typically encounter. Firstly, when a parent goal becomes active, all subgoals (and their related actions) become active as well. However, it is seldom the case that a user is simultaneously aware of current and future goals. Rather, user behaviour is typically more anticipatory and reflexive (Endsley, 2000; Suchman, 1987). According to HTA, subgoals and their actions are maintained as conscious constructs available to the user. Again, this is unlikely to be the case, as many procedural actions essential for effective task completion are habituated or unconscious even when the associated goal is being actioned (Whittaker et al., 2000).
GroupWare Task Analysis

Developed by van der Veer, Lenting, and Bergevoet (1996), GroupWare task analysis (GTA) attempts to incorporate cooperative strategies within cognitive task models. Within GTA, complex tasks are decomposed into low-level unit tasks (Limbourg & Vanderdonckt, 2004). However, by assigning roles to both tasks and agents in the environment, GTA differs from HTA in its ability to model aspects of collaboration and cooperation. As roles change within the organization and across activities, traits that link specific tasks to activity structure, agents, and artefacts can be identified. This process contributes a social dimension to task performance analysis. Unfortunately, GTA currently lacks sufficient formalism, and therefore expressive power, to handle complex environments. Additionally, GTA is presently only effective when evaluating existing systems. It is therefore not appropriate for exploratory modelling activities.

Cognitive Task Analysis

Cognitive task analysis (CTA) refers to a host of techniques that can be applied to investigating expert system usage (Gordon & Gill, 1997). Examples include naturalistic decision making, recognition-primed decisions, and schema theory (Zsambok & Klein, 1997). CTA techniques address many of the weakness inherent to HTA and GOMS by explicitly modelling procedural knowledge and complex information-processing structures (Hollnagel, 2003). Decision ladders capture the central characteristic of information processing, yet incorporate short circuits, or experience-based patterns of action that permit experienced individuals to move between goal structures in response to changing action and planning resources. CTA is therefore highly suited to modelling expert behaviour. However, analysing sociocultural features of an activity is not sufficiently developed or formalised in CTA. Although HTA and GOMS may be cast as derivatives of the CTA approach, CTA is inherently more adaptable, as seen through (1) its advocacy of methodological plurality; (2) promotion of task heterarchies over hierarchies; and (3) its consideration of both cognitive and environmental artefacts within an activity. As CTA models activities in terms of cognitive architectures, it is less applicable to investigating exploratory, descriptive, and ill-structured social activities.

Cognitive Work Analysis

CWA is a work-centred framework that represents a major advancement upon traditional CTA approaches. It is designed specifically for analysing work environments that require significant human behavioural adaptation to new and existing tasks (Rasmussen et al., 1994; Vicente, 1999; 2004). It is built upon a meta-cognitive task model that integrates both cognitive and ecological practices within a holistic framework for work practice analysis.

CWA differs from other cognitive approaches in that it does not merely focus on normative modelling (how HCI should proceed); descriptive modelling (how HCI currently occurs); or formative modelling (how specific actions are sequenced during an activity). Rather, the aim of CWA is three fold:

1. Identify the properties of the work environment
2. Define likely work boundaries
3. Evaluate boundary effects on HCI

In CWA, actor-system interaction is always situated within some activity; hence, it considers not just the tasks actors perform, but also the situation in which they are undertaken. In conducting a work analysis, activity is broken down into five levels of analysis: (1) work domain analysis; (2) control task analysis; (3) strategies analysis; (4) social-organisational analysis; and
(5) work competencies analysis. In practice, these five levels are used to constrain possible human-computer design considerations. Although prescriptive methods are formalised and integrated within CWA, the framework itself advocates a plurality of analysis techniques, so long as they fulfil a practical purpose. As conceptual levels are interdependent (soft boundaries), analysis progresses from one layer to another in an iterative fashion — all layers activated to some lesser or greater extent at any point in time. Moreover, in applying CWA, there is no explicit ordering of analysis activities (Vicente, 1999). This renders the approach extremely flexible and reduces the learning curve required to apply its principles in practice.

Work domain analysis seeks to establish the structural limitations of the work domain. This stage identifies the type of users expected to participate in the activity, the activity problem, and additional logical and physical constraints upon the system. Control task analysis represents the set of abstract tasks that control the general flow of information within an activity. Strategies analysis investigates how individuals or groups overcome a problem when presented with a set of alternatives. This decision-making process is similar to that of selection rules in GOMS. However, unlike GOMS, this process considers prior experiences, spatio-temporal constraints upon operation, motivation, aptitude, and other similar criteria. Social-organisational analysis seeks to establish how tasks are shared between actors and artefacts in a cooperative and collaborative manner. Finally, work competencies analysis examines the degree of training and experience actors bear upon some activity.

Summary of Approaches

To determine which approach may best be used within a particular Web design project, Table 4 provides a brief overview of the cognitive approaches previously mentioned and demonstrates their shared and unique attributes.

Ecological Task Analysis

Until recently, task analysis in HCI has focused almost exclusively upon deriving reductionist cognitive representations of individual tasks (Hollan, Hutchins, & Kirsch, 2001). Although HCI is

<table>
<thead>
<tr>
<th>Approaches</th>
<th>GOMS</th>
<th>HTA</th>
<th>GTA</th>
<th>CTA</th>
<th>CWA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expressive power</td>
<td>LOW</td>
<td>LOW</td>
<td>MEDIUM</td>
<td>HIGH</td>
<td>HIGH</td>
</tr>
<tr>
<td>Complexity</td>
<td>LOW</td>
<td>LOW</td>
<td>MEDIUM</td>
<td>HIGH</td>
<td>HIGH</td>
</tr>
<tr>
<td>Collaboration</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Timing</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Roles and responsibilities</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Evaluation</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Requires Training</td>
<td>LOW</td>
<td>LOW-MED</td>
<td>MED-HIGH</td>
<td>HIGH</td>
<td>HIGH</td>
</tr>
<tr>
<td>Scalable</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Social Orientation</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
</tbody>
</table>
currently moving towards more ecologically valid models of human praxis, postcognitivist theorists have yet to agree upon the most promising alternatives (Kaptelinin, 1996a). Resolving this issue depends on two factors:

1. Establishing a common unit of analysis
2. Constructing integrative frameworks that model both cognitive and social phenomena

The need for integrative frameworks is especially relevant to Web designers who must not only support existing social practices through a novel medium, but also the cognitive demands imposed by semiotic constraints when delivering information via the Web (Smart, Rice, & Wood, 2000). Table 5 outlines the required properties of an integrative framework.

At the heart of the postcognitive movement lies the notion of situation awareness. Situated awareness represents an understanding of social norms and rules, or a state of knowledge about the context, situation, and environment in which an activity takes place (Erickson & Kellogg, 2001). Situation awareness does not describe, per se, the process of acquiring knowledge, but rather is the end result of invoking any number of cognitive and meta-cognitive processes during active participation in a practice (Endsley, 2000). Furthermore, situation awareness is distributed among individuals, and hence, actionable and relevant only within specific communities of practice (Gasson, 1999). Situation awareness requires individuals to maintain an awareness of the continuously unfolding characteristics of an activity. Summarising Endsley (2000), situation awareness entails:

Table 5. Postcognitive task analysis criteria

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formalised notation system</td>
<td>Common, accessible notation that facilitates conceptual modelling and communication of knowledge between interested parties (Balbo et al., 2004; Erickson &amp; Kellogg, 2001)</td>
</tr>
<tr>
<td>Methodological flexibility</td>
<td>The ability to adopt existing methodologies rather than constantly reinvent the wheel (Whittaker et al., 2000)</td>
</tr>
<tr>
<td>Cost-effective practices</td>
<td>Cost-effective practices that encourage the use task analysis during system design, thus increasing return-on-investment (Stanton, 2004)</td>
</tr>
<tr>
<td>Expressive power</td>
<td>Practical methodologies with sufficient expressive power to support novice and expert usage (Balbo et al., 2004)</td>
</tr>
<tr>
<td>Integration within formal software engineering frameworks</td>
<td>An ability to be applied across the SDLC, including integrative mechanisms that facilitate software engineering practices (Farmer, 2006; Farmer et al., 2004; Flor &amp; Hutchins, 1991; Kazman, Abowd, Bass, &amp; Clements; 1996; Rosson &amp; Carroll, 2002)</td>
</tr>
<tr>
<td>Reuse</td>
<td>The ability to support that HCI claims reuse (Carroll, 1996; Sutcliffe, 2000)</td>
</tr>
<tr>
<td>Ecological validity</td>
<td>Research findings and technology must support existing social processes and cultural practices (Nardi, 1996)</td>
</tr>
</tbody>
</table>
1. Perceiving relevant contextual information
2. Understanding how the roles and relationships between actors impacts information presentation and communication
3. An ability to use social cues to assist in planning for future events

Albers (1998) states that the quality of an individual’s situation awareness is governed by their ability to comprehend, integrate, and reconstruct new information. This three-step process involves extracting the relevant basic information from the task/activity, integrating the information into the individual’s conceptual understanding of the problem, subsequently generating new rules and beliefs so as to extrapolate the information into future problem-solving situations. As situations change and new information is presented, so must the individual’s awareness of emerging conditions that are likely to impact upon their effective participation in an activity. The problem of situation awareness is highly relevant to developers of Enterprise Content Management systems where the most critical factor impacting successful deployment and integration is user adoption. Postcognitive task analysis frameworks that maintain the centrality of situation awareness are more likely to succeed at meeting this challenge.

Goals, experiences, expectations, and motivations are interdependent constructs that produce localised expectations about situated action, and as such influence situation awareness (Endsley, 2000). Goals help to orientate an individual across a series of tasks, including error recovery. For example, when a user encounters an error during interaction with a Web-based form, a breakdown in planning occurs. In assessing available feedback, the user must evaluate available actions against their goals and state of knowledge about the activity. In overcoming the breakdown or conflict, emergent behaviour may not only result in a reprioritisation of individual goals, but may also lead to new interaction strategies.

Expectations are influential in the division of attention during HCI, acting as a selection mechanism, reducing demand on cognitive resources to efficiently respond to new information. Experience is also influential in the development of automaticity. For instance, expert system users are expected to demonstrate better skills compared to novices, as they are likely to have developed more efficient strategies and greater responsiveness to overcoming breakdowns. The difficulties novices face may be a result of their lack of habituated actions during HCI, rendering the task more cognitively demanding. Situation awareness among novice users typically increases through gradual exposure to, and experience with, the system. However, increased situation awareness does not imply good task performance. It is a probabilistic construct, making good performance more likely, yet not assured.

The notion of situation awareness is found within several postcognitive HCI frameworks. Here, we present a broad comparison of a few notable frameworks. We examine how each of the competing frameworks conceives of an appropriate unit of analysis, and how this unit of analysis impacts each framework.

**Situated Action Models**

The situated action view of HCI treats knowledge as a subjective construct. Knowledge is inextricably linked to interpersonal communication and contextual activity. Knowledge is not only subjective, but specifically relevant to a particular situation in which some work practice occurs (Gasson, 1999). Nardi (1996) notes that situated action perspectives are more focused on context-sensitive practices than on the cognitive properties of the artefacts with which individuals interact.

The unit of analysis in situated action is the motivated activity within a relevant community
of practice (Lave, 1988), supplanting the largely cognitivist unit of analysis, task. Within each of the ecological perspectives treated in this chapter, the notion of task is redefined or appropriated to mean action. Subjugating the importance of task to activity promotes the influence of actor roles, relationships, and sociocultural cues within a particular setting.

One of the key characteristics of the situated action model is its adherence to reflexivity in system design. Experience, cultural and historic beliefs and anticipatory planning on behalf of the individual are therefore de-emphasised within this perspective. Rather, situated action focuses on the immediacy of interaction, the fluctuating conditions of the situation, and the learning opportunities provided by an individual’s creative response(s) to alternative paths through the activity (Lave, 1988; Nardi, 1996; Suchman, 1987). An activity’s structure is not something that precedes interaction but emerges directly out of the immediacy of the situation. Planning and goals are produced in the course of action, and thus are work products, rather than characteristics that orient the activity (Bardram, 1998). This view implies that people are basically opportunistic. Planning and goals either emerge concurrently out of participation in the activity itself, or by conscious reflection upon previous interaction.

As situated action subscribes to the view that analysts should only be concerned with recording observable behaviour (Endsley, 2000), we believe that it is not appropriate for analysing exploratory domains, where it is crucial to first identify the contextual problem under investigation and subsequently determine the expected outcomes of the activity. Similarly, we feel that the approach is not sufficiently descriptive to examine large collaborative work activities where knowledge and planning are often spread across actors. In large organisations, activity is seldom seen as a knee-jerk reaction to current events.

Distributed Cognition

Distributed cognition is concerned with how knowledge is propagated and transformed by agents within activity. An agent is any cognitive artefact of the system, be it human, machine, or other work product. The unit of analysis is the cognitive system. Distributed cognition relaxes the assumption that the individual is the best or only useful unit of analysis and thus extends the reach of what is considered cognitive to both systems that are smaller and larger than the individual (Hollan et al., 2001; Hutchins, 1995). The cognitive system in distributed cognition is thus more akin to the term complex cognitive system.

Goals, according to distributed cognition, are not merely maintained within the mind of the subject (individual or group), but rather embedded within the cognitive system. Distributed cognition posits that artefacts may themselves possess goals. The cognitive system can only be understood when we know the contributions of individual agents, their shared contributions and collaboration strategies, and the nature agent behaviour in the environment. In contrast to situated action, distributed cognition incorporates culture, context and history, but from within an embedded perspective (Hollan et al., 2001).

There are striking similarities between the distributed cognition and activity theory (Nardi, 1996). Both are activity-centric: they recognise activity as a hierarchical, goal-oriented structure; align physical, verbal, and nonverbal actions with specific goals; and they distinguish between conscious and unconscious actions. Additionally, neither framework prescribes a particular set of methodological practices. Nevertheless, there are two notable differences between the approaches:

- Activity theory is essentially human-centric, individuals motivating activity.
• In activity theory, artefacts are not seen as goal-oriented, cogent entities.

In contrast to activity theory, distributed cognition holds that in cooperative environments, individuals maintain only partial models of the problem. Hence, for distributed cognition, complex systems rather than individuals are the appropriate unit of analysis.

Latour (1987) provides theoretical support for this last argument, stating that part of what is technological is social, and what is social is technological. As relationships in actor-networks constitute human-human and human-artefact interaction, a change in content structure, user interface, pedagogy, or number of participants in an activity has the potential to change the nature of an activity. While not implying that an artefact is cogent, actor-network theory suggests how a work product may impact social and cognitive processes (Tatnall, 2002).

Activity Theory

Activity theory is a descriptive conceptual framework that has emerged primarily from contributions by Vygotsky (1986), Leont’ev (1978), and Engeström (1987). Activity theory serves to describe the different forms of human praxis and developmental processes involved in HCI. Activity theory represents a truly ecological approach to task analysis, providing a broad theoretical framework for examining collaborative activities (Stanton, 2004). As the name suggests, the unit of analysis in activity theory is the activity itself. Individual and social processes are interwoven in the generation of contextualised meaning. Similar to distributed cognition, individual knowledge and meaning cannot be separated from the context of an activity.

Activity theory (depicted in Figure 3) is an object-oriented, task analysis framework. The importance of goals is present, however they are subordinate to motivated outcomes, or objective purpose (object) (Whittaker et al., 2000). Unlike situated action, interaction is not opportunistic, but purposeful. Purpose is context specific as activities define context, and context defines an activity. Therefore, the act of doing in one context cannot be considered congruous with the act of doing in another (Lave & Wegner, 1991). This is in direct contrast to cognitive task models such as GOMS.

One of the principal tenets of activity theory is tool mediation. Activities are mediated by individuals carrying out their tasks through the use of available artefacts in the environment. When an artefact is used to mediate interaction, it is said to become a tool. A tool can be any material or mental process used to transform or convey

Figure 3. Activity theory framework
Understanding the Nature of Task Analysis in Web Design

information (Kaptelinin, 1996b). As tools exist only through purposeful activity, they contain cultural and historical residue. They are therefore capable of imparting external sociocultural beliefs and historic conventions upon an activity.

Activity theory posits a hierarchical structure of activity, consisting of three primary elements: activity, object-oriented actions, and operations (Leont’ev, 1978). As shown in Figure 3, activity is motivated by the subject (individual or group). An activity is comprised of mediated, object-oriented actions. Each action is associated with one or more parent and subgoals. This implies that actions are polymotivated; each action (task) may be applied to accomplish several goals. This is important as it provides a way of supporting alternative paths or strategies within an activity for obtaining the purposeful outcome. Accordingly, we cannot focus purely on the action itself, as its purpose or role is ambiguous. Lastly, operations are habitual behaviours or unconscious actions.

According to activity theory, there is constant interaction between these three elements of human praxis. This dynamicism is critical to our understanding of human behaviour in interactive environments. However, despite the advantages of applying an activity theoretic approach to interactive system design, there is as yet no systematic way of interfacing the approach with rigorous software engineering methodologies (Turner, Turner, & Horton, 1999).

Scenario-Based Design

Scenario-based design (SBD) promotes the use of scenarios (or structured narratives) in HCI as a way of understanding human activity and creating computer systems (Carroll, 2000). Today, they are widely used across disciplines. In an attempt to render software development more social, Carroll (1996, 1997) argues that activity theory can be applied effectively to SBD. While Carroll does not describe how this can be accomplished, suggestions are found elsewhere (Carroll, 2000; Go & Carroll, 2004; Kazman et al., 1996; Rosson & Carroll, 2002). Admittedly, natural language narratives are not the most scientific notation used within HCI or software engineering; however, it is often by these means that technical information is conveyed throughout the development effort (Carroll, 1996). Moreover, scenarios provide a flexible mechanism for integrating “real life” accounts of activity with the more empirically discrete views employed within software engineering.

During discovery and definition (see Table 1), scenarios can be used speculatively at the start of the SDLC to document expected user behaviour or to describe hypothetical activities. In the absence of an existing system, this process improves ecological validity during requirements engineering (Stanton, 2004). To assist with requirements engineering, Kaptelinin et al. (1999) suggest that Activity Checklists can assist practitioners to focus on salient aspects of an activity, thereby constraining the process of requirements engineering during SBD.

Scenarios represent purposeful interaction within a system, and are inherently goal-oriented. Scenarios are always particular to a specific situation, and thus situated within a particular frame of research. Because scenarios are actor-driven, describing interaction from the perspective of at least one individual, they can be effectively integrated with an activity theory framework. Finally, as scenarios use a common notation, natural language, all stakeholders in the development process can easily communicate requirements, experiences, and other opinions/beliefs without requiring extensive training. Consequently, scenarios serve as an excellent lingua franca for communication between all project stakeholders; a primary goal of any good task analysis framework (Balbo et al., 2004).
Table 6. Overview of postcognitive task models

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Situated Action</th>
<th>Distributed Cognition</th>
<th>Activity Theory</th>
<th>Scenario-Based Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expressive power</td>
<td>LOW</td>
<td>LOW</td>
<td>MEDIUM</td>
<td>HIGH</td>
</tr>
<tr>
<td>Complexity</td>
<td>LOW</td>
<td>LOW</td>
<td>MEDIUM</td>
<td>HIGH</td>
</tr>
<tr>
<td>Collaboration</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Timing</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Roles and responsibilities</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Evaluation</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Requires training</td>
<td>LOW</td>
<td>LOW-MED</td>
<td>MED-HIGH</td>
<td>HIGH</td>
</tr>
<tr>
<td>Scalable</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Social orientation</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>

Summary of Approaches

Table 6 provides a brief overview of the postcognitive approaches described previously and demonstrates their shared and unique attributes.

TOWARDS SITUATED TASK ANALYSIS IN WEB DESIGN

In this chapter we have outlined the limitations associated with both cognitive and postcognitive approaches to task analysis. Cognitive approaches suffer from their inability to model dynamic social processes; to consider aspects of interaction which are not reducible to discrete cognitive events; and to elicit and describe user goals and their limited extensibility and transferability to new domains. Postcognitive approaches on the other hand suffer from a lack of tools and techniques that bridge theoretical frameworks and practical design and evaluation methods. Although promoting increased reuse of HCI knowledge, postcognitive frameworks are less computationally powered, thus possessing a significant barrier to their integration into standard Web design practices as described in Vora (1998). The challenge therefore is to promote integrative, task analysis frameworks that support both cognitive and ecological perspectives (Czerwinski & Larson, 2002). This new breed of task analysis frameworks is referred to as situated task analysis (Farmer & Hughes, 2005b).

Situated task analysis incorporates the view that in natural settings, problems are frequently ill defined, requiring communication, collaboration, negotiation, and prioritisation. This broadens the field of regard from the individual to include organisational and social contexts. This situated view can be seen as an example of macrocognitive architecture, describing cognitive functions as they are performed in natural settings (Klein,
Understanding the Nature of Task Analysis in Web Design

Ross, Moon, Klein, Hoffman, & Hollnagel, 2003). This is in opposition to the traditional microcognitive view of task analysis that focuses on the building blocks of cognition. A detailed account of a situated task analysis framework: cognition, activity, social organisation, environment (CASE) is provided in Farmer (2006).

A Simple Approach to Situated Task Analysis in Web Design

A major obstacle in developing Web sites that are both usable and fit-for-purpose is capturing stakeholder requirements and communicating them to other project team members, especially designers. User requirements are often qualitatively produced, resulting in rich descriptions of functional and operational requirements. While this assists with understanding the nature of activity, it does not always suit developers who are required to describe the system in terms of discrete, verifiable behaviour. In the following, we describe an iterative five-step process that facilitates communication and integration of both static and dynamic views within the requirements engineering and design phases of Web site development.

Step One: Requirements Engineering and Goal Generation

The first step in our process is requirements engineering. Scenarios are ideal in this situation as they are written in natural language and are therefore highly accessible and require little practitioner training (Go & Carroll, 2004).

Figure 4 describes the iterative process of eliciting user requirements using activity checklists (Kaptelinin et al., 1999) and scenarios. Scenario descriptions are iteratively refined until they are stable, the outcome being a set of scenarios reflecting multiple viewpoints of interaction. Using an initial Activity Checklist, containing likely questions and statements related to the purpose of the activity, and a minimal-use scenario, we initiate the requirements elicitation process. As shown in Figure 4, our first aim is to elicit initial traits or themes that emerge out of the user’s perception of the activity. Users may be additionally asked to develop their own use scenario to increase participation and contextuality during the development process. This process continues, redefining the scenarios and checklists, until we are able to group these traits into sets of themes or categories. At each stage, the new scenarios are...
validated against the users themselves in order to maintain internal consistency.

In a similar manner, once no further traits emerge from our analysis, we proceed by reducing these initial categories to a core set of salient concepts. These concepts can be considered high-level objectives within the activity or goals. These concept-goals can be considered fairly stable and transferable between users, as they will have been derived from multiple perspectives. Finally, we arrive at a set of scenarios that represent the key concept-goals of our activity. These scenarios can be linked together to provide a chain or hierarchical view of the scenarios relative to a particular requirement. The corollary of this process becomes increasingly important as we move towards the initial design phase.

In Figure 5 we see more clearly this process of scenario refinement and linkage. Our initial minimal use scenario (R1.1) is iteratively refined to reveal particular traits of the activity (R1.2). As this ultimately is an exploratory process, each stage can be seen as a form of goal refinement. We notice in Figure 5 that within this process of refinement, conflicts between scenario views (i.e., between users or between users and developers) result in an update of the Activity Checklist. In this way, we are able to capture the context of use and apply this knowledge to subsequent elicitation cycles. As we move towards concept-based scenarios in our analysis, we move away from action-oriented perspectives towards more holistic, goal-oriented views of activity.

**Step Two: Determine Salient Features of Interaction**

Having constructed a core set of requirements containing both low and high-level views of interaction, the next step is to determine which concepts are most critical to the success of the Web application. Although we already possess relatively stable concept-goals with which to initiate our design process, some concepts may be more salient than others. Techniques for determining salient cognitive and social features during interaction are described in Farmer (2006).

*Figure 5. Iterative goal-driven scenario development*
Step Three: From Scenarios to UML Use Cases

Upon establishing the most salient concept-goals with the system, the next step is to transform these rich textual descriptions into static, structural views, amenable to inclusion in the design process. This can be achieved using unified modeling language (UML), use case diagram notation (Rumbaugh, Jacobson, & Booch, 1999). We start by modelling the high-level use cases. These typically correspond to the concept-goal scenarios previously elicited, and hence provide a goal-oriented system view. Subsequently, we drill down into each of these use cases, decomposing them into their subsystems and determining their structural constraints.

Use case decomposition, as shown in Figure 6, is the process of iteratively decomposing an activity into its constituent parts. This process has already been achieved via the bottom-up process of concept-goal generation. Therefore, use case decomposition may easily be achieved by associating each concept-goal with its related categories and traits. This information is present within the previously constructed scenario chains. Designers are therefore able to isolate requirements at particular levels of granularity as well as investigate the structural relationships between use cases (implements, depends, generalises) via the rich descriptions provided in the scenarios.

Figure 6 provides an example of a use case decomposition for an online shopping activity. In this instance, the user has only four concept-goals. These are search catalogue, view shopping cart, buy book, and change profile. Common concepts that are shared among requirements, such as login, can be modelled via a dependency relationship. Modelling the relationship between a concept and a category (see Figure 4) can be achieved through inheritance, or implementation. In addition to modelling relationships, designers may choose to isolate particular concept areas, such as social organisation, to further narrow the design perspective.
**Step Four: From Use Cases to Design**

After constructing the UML use cases at various levels of granularity (from goals to actions), the next step involves transforming these formal decompositions into design elements. One way to manage this complex process is to establish distinct viewpoints, modelling the system from various perspectives (Rumbaugh et al., 1999). Here, we only consider the structural and behavioural system views. The major distinction between these views is that the former emphasises object-oriented interaction, whereas the latter describes event-driven interaction.

The structural view provides a physical description of how the Web site will be made up of various components, including relationships between those components. This is primarily achieved using UML class diagrams that provide an object-oriented view of the system. While primarily promoting a static view of the system, object methods and attributes can also be shown. Using the use cases generated previously from the concept scenarios, it is possible to rapidly construct a high-level representation of the site structure, objects acting as placeholders for each concept goals.

Having modelled at a conceptual level the key user requirements as UML classes or objects, we use the link structure provided during the initial requirements engineering phase to determine additional objects, methods, and attributes of the system. For instance, we can model category-based scenarios in UML class notation via aggregate relations (“is-part-of”) to our concept-based objects (Figure 7). Additional techniques, such as noun extraction (noun=object, verb=method, adjective/adverb=attribute), may be used to extract structural information from the scenarios themselves (Schach, 1999).

Figure 7 provides a simple demonstration of this process. In this instance, the class objects **Login** and **Search** represent complex objects associated with login management and searching a Web site. We note these relationships could have been modelled alternatively as methods (login and search are verbs) within other classes, or as interfaces that could be implemented by other objects, as shown in Figure 6.

This process can be extended to the trait-based scenarios to establish additional object attributes.

---

*Figure 7. UML Class diagram representation of concepts*
This process has several benefits: it provides a rapid means of determining initial site structure within a goal-oriented framework; it permits designers to link design decisions back to initial requirements, as well as examining requirements from various stakeholder points of view and at various levels of granularity; and designers may link additional design-centric scenarios to each requirement, providing details such as event sequencing and information flow, therefore improving traceability between the SDLC phases.

In examining methods and attributes more closely using our scenarios, we are likely to become more focused on event-driven behaviour. In developing a behavioural system view, we are interested in modelling the flow of system information through event or function-driven interaction. We typically see functional flows (Appendix A) and sequence diagrams used at this point in the initial design phase. These techniques constrain interaction to a finite-state representation, limiting users to predetermined paths through the system. Where multiple paths may exist, functional flow analysis and task models may be used to determine optimal paths through the system in order to reduce variability and improve site effectiveness and efficiency. At this point, our design analysis has moved away from modelling goals to modelling actions and evaluating task performance.

Finally, these event-driven views may be abstracted to form UML dialog models or activity diagrams, providing a means of aggregating task models or flow analyses. These dialog models may then be linked into our initial scenario/requirement chain to provide additional system information. This process ensures a multi-view representation of requirements that are linked forwards and backwards to design. Moreover, it ensures that both goal-oriented and action-oriented perspectives of interaction are conveyed during development.

**Step Five: Start Over**

We recall that task analysis is not a once-off process (Balbo et al., 2004). This is especially true of interactive systems and environments, such as the Web. It is important to note that most people actually learn by doing (Lave & Wegner, 1991). Rather than thinking of a site’s design as merely being usable, we should consider how the site or application, as a tool, facilitates participation in an activity.

Developing effective Web sites requires consideration of both novice and expert users. Most users, however, do not arrive at a site as experts. They reason, use past experiences, and explore through interaction. Their ability to gain expertise in these virtual environments is partially governed by how well the design supports and extends the user’s current capabilities. This can only be achieved when there is sufficient coherence between the context of use and the virtual environment’s design. Taking a user-centred approach to Web design, including an appropriate task analysis framework and set of methodological practices, ensures that developers will never be too far away from their users. Ultimately, it is they who will judge the relative merits of your site/application.

**CONCLUSION**

Web design is a complex activity at the best of times. Not only do designers frequently encounter technological limitations imposed by a novel communication medium, but also they are highly isolated from their users. Moreover, arguably more than with any other technological medium, the Web application target audience is extremely heterogeneous. Therefore, the ambiguous and diverse nature of Web application use imposes critical limitations on Web design practices. Shadowed by the importance of developing
quality, fit-for-purpose systems, this chapter has highlighted the considerable benefits to be gained by incorporating task analysis in Web design. Specifically, this chapter has described how task analysis is not a single process that can be applied indiscriminately of context and use. It is a highly stylised and domain-dependent activity that occasionally suffers from conflicting theoretical and methodological approaches.

We have argued that task analysis is not a silver bullet, rather a means of “getting to know” your users and communicating this information through product development and evaluation. Rather than describing a core set of prescriptive guidelines for conducting task analysis in Web design, this chapter has set out to inform the reader via a critical examination and principled discussion of various approaches to task analysis. Consequently, our hope is that the reader as practitioner will be both more aware and appreciative of the role task analysis can and should play in Web design.

**ACKNOWLEDGMENTS**

We would like to thank Baden Hughes and Sandrine Balbo for their collaboration and comments on related work. Special thanks go to Sandrine Balbo for her presentation on task models in HCI, which spawned an insightful and lively afternoon of discussions that stimulated aspects of our discussion here.

**REFERENCES**


Understanding the Nature of Task Analysis in Web Design


McDonald, Y. Waern, & G. Cockton (Eds.), *People and computers XIV* (pp. 3-21). Bath University, UK: Springer-Verlag.

APPENDIX

Task analysis matrix adapted from a survey conducted by Bonaceto and Burns (2003).

<table>
<thead>
<tr>
<th>Key (applicability)</th>
<th>Category</th>
<th>Method</th>
<th>Task Design</th>
<th>Interface Development</th>
<th>Workload Estimation</th>
<th>Training Development</th>
<th>Problem Investigation</th>
</tr>
</thead>
</table>

This work was previously published in Human Computer Interaction Research in Web Design and Evaluation, edited by P. Zaphiris, pp. 78-107, copyright 2007 by Information Science Publishing (an imprint of IGI Global).
Chapter 2.31

Designing for Tasks in Ubiquitous Computing: Challenges and Considerations

Stephen Kimani
Jomo Kenyatta University of Agriculture and Technology, Kenya

Silvia Gabrielli
University of Rome “La Sapienza”, Italy

Tiziana Catarci
University of Rome “La Sapienza”, Italy

Alan Dix
Lancaster University, UK

ABSTRACT

The traditional desktop computing paradigm has had major successes. It also should be noted that we are in a day and age where many good computer and device users are increasingly finding themselves being required to perform their activities not in offices/desktops but in real-world settings. Ubiquitous computing can make possible in the real-world setting what would have otherwise been impossible through desktop computing. However, there is a world of difference between the real-world and the desktop settings. The move from the desktop to the real-world settings raises various issues when we consider the nature of tasks that the ubiquitous devices/applications would be expected to support and the real-world context in which they will be used. A careful study of the nature of tasks in ubiquitous computing can make some design requirements in the development of ubiquitous applications more evident. This chapter proposes ubiquitous application design and evaluation considerations emerging from a deeper understanding of the nature of tasks in ubiquitous computing.
INTRODUCTION

It is worth acknowledging that the traditional desktop computing paradigm has had major successes. On the same note, it should be observed that we are in a day and age where many people have become good computer and device users. However, these users are increasingly finding themselves performing or being required to (or having to) perform their activities not in offices and desktops but in the real world settings. In describing the situation, Kristoffersen and Ljungberg indicate that the hands of such users “are often used to manipulate physical objects, as opposed to users in the traditional office setting, whose hands are safely and ergonomically placed on the keyboard.” (Kristoffersen & Ljungberg, 1999). It is interesting to observe how ubiquitous computing can come in handy toward making possible in the natural setting what would have otherwise been impossible through the desktop computing paradigm. It is therefore not uncommon to encounter a user who “carries out one or many parallel activities from virtually anywhere at anytime while at the same time interacting with other user(s) and/or device(s).” (Bertini et al., 2003).

However, it is worth noting that there is a world of difference between the real world setting and the desktop setting. As we consider the move from desktop computing (fixed user interfaces) to the real world settings, various issues and demands arise when we consider the nature of tasks the ubiquitous devices/applications (and thus ubiquitous user interfaces) would be expected to support and the real world context in which they will be used.

Consequently, it does turn out that a careful study of the nature of tasks in ubiquitous computing can make some requirements in the design and evaluation of ubiquitous applications become more evident, which forms the basis of this chapter. In particular, we will describe the nature of tasks in ubiquitous computing, and then propose and describe ubiquitous application user interface design and evaluation considerations emerging from a deeper understanding of the nature of tasks in ubiquitous computing.

The rest of the chapter is organized as follows; it first provides some background knowledge. It then gives an overview of the nature of tasks in ubiquitous computing. After that we propose and describe ubiquitous application design and evaluation considerations respectively based on the foregoing. We then highlight some open issues and conclude the chapter.

BACKGROUND KNOWLEDGE

In this section, we describe some of the key concepts relevant to the chapter. In particular, we describe ubiquitous computing. It should be noted that in the history of computing, the requirement to take into consideration the real world context has arguably never been more critical and pressing than in this day and age of ubiquitous computing. After describing ubiquitous computing, we then focus the description on the concept of context.

Ubiquitous Computing

Weiser coined the term ubiquitous computing (ubicomp) and gave a vision of people and environments augmented with computational resources that provide information and services when and where desired (Weiser, 1991). Dix et al. define ubicomp as: “Any computing activity that permits human interaction away from a single workstation” (Dix et al., 2004). Since then, there have been tremendous advances in mobile and wireless technologies toward supporting the envisioned ubiquitous and continuous computation and, consequently, ubiquitous applications that are intended to exploit the foregoing technologies have emerged and are constantly pervading our life. Abowd et al. in (Abowd et al., 2000) observe
that ubicomp applications are characterized by the following:

- **Natural interfaces**: Supporting interaction techniques that permit humans to interact with computing machines through the use of more natural interaction paradigms (e.g., speech, gesture, pen writing).
- **Context-awareness**: Ubicomp applications are expected to exploit the whole set of computing and telecommunication technologies that operate taking into account the context.
- **Automatic capture of live experiences**: Ubicomp applications often adopt or provide techniques that enable the user to record elements of their live experiences (e.g., photos, video, audio) and the management of the same.

**Context**

Context has been defined as “any information that can be used to characterize the situation of an entity.” (Dey, 2000), where an entity refers to “a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.” (Dey, 2000). Context entails aspects such as location, infrastructure, user, environment, entities, and time. The infrastructure could include technical resources such as server and network capabilities and connections, applications, and so forth. User includes user data/profile, usage patterns, and so forth. The environment refers to the physical condition of the setting an could include light, temperature, and so on. Entities refer to people, devices and objects. Time could include date, time of the day, season, and so on. Abowd et al. provide in (Abowd et al., 2000) a review of ubicomp research and summarize context in the form of “five W’s”:

- **Who**: As human beings, we tailor our activities and recall events from the past based on the presence of other people.
- **What**: Perceiving and interpreting human activity is a difficult problem. Nevertheless, interaction with continuously worn, context-driven devices will likely need to incorporate interpretations of human activity to be able to provide useful information.
- **Where**: In many ways, the “where” component of context has been explored more than the others. Of particular interest is coupling notions of “where” with other contextual information, such as “when.”
- **When**: With the exception of using time as an index into a captured record or summarizing how long a person has been at a particular location, most context-driven applications are unaware of the passage of time. Of particular interest is the understanding of relative changes in time as an aid for interpreting human activity. Additionally, when a baseline of behavior can be established, action that violates a perceived pattern would be of particular interest.
- **Why**: Even more challenging than perceiving “what” a person is doing is understanding “why” that person is doing it.

**NATURE OF TASKS**

The interaction of the user with the ubiquitous device/application can be viewed in at least two dimensions:

- user-ubiquitous application interaction dimension and
- user-ubiquitous device dimension.

User-ubiquitous application interaction dimension entails tasks in which the user is primarily interacting with the ubiquitous application and
the device I/O modalities in order to access services such as support services (e.g., emergencies, help/service), information services (e.g., gathering/recording information, accessing/retrieving information, sharing information, communicating) and entertainment services (e.g., games, music, videos). User-ubiquitous device dimension categorizes tasks that entail the actual handling of the device (such as holding the device, wearing the device, attending to the device).

There are situations whereby interaction with the ubiquitous application, though important, is not the primary task but rather a secondary/supplementary task. In such a case, such ubiquitous devices/applications would be used to provide support/assistance and gather/make available some resources (such as information) on behalf of a user who is engaged in another primary task in the real environment/setting. In fact, the tension between the primary tasks of a ubiquitous user and the user’s interaction with the ubiquitous device/application can be seen in the literature (e.g., Pascoe et al., 2000). Notwithstanding the foregoing, there also are situations in which interaction with the ubiquitous application is the primary contributor to the user’s accomplishment of the primary task; interacting with the ubiquitous device/application can be viewed as directly carrying out the primary task. In this case, the use of the ubiquitous device/application tends to be more intimately connected with what the user is really doing (or intends to achieve) in his/her embodied/physical self. For instance, where the user is using the ubiquitous application to inform him/her about the location he/she is in. However, it is also worth pointing out that at different time granularities, the primary task and secondary task may swap in ubiquitous computing. The foregoing situations raise challenges and that would need to be taken into consideration when designing and evaluating (developing) ubiquitous application user interfaces.

Some ubiquitous interactions are low-intention or incidental interactions—that is one where the user is not focused on the interaction being supported by the ubiquitous system and may not even be directly aware of the support (Dix et al., 2004, Dix, 2002). In such cases, the primary task or purposeful task needs to be understood in order to interpret sensor data. However, the supported task also needs to be understood in order to be able to administer appropriate support. This distinction is likely to be useful in other context sensitive situations. For example, if a user has recently consulted an online diary related to a coming birthday and a short while later starts to initiate a phone call, then it may be appropriate for the number of the person whose birthday is coming to be at the top of the list of suggested numbers. The former task, looking up the diary entry, is the initial purposeful task and needs to be interpreted in order to determine which people are important or relevant. The latter task, phoning, is the supported task and the knowledge of which people are currently significant is used to support this task.

While simpler task models assume a level of pre-planning or training, it is expected that most ubiquitous interactions are worked out at the moment (see situated action below) or maybe triggered from things in the environment (Dix et al., 2004b). Furthermore, being in a particular place or meeting a particular person may prompt tasks or activities that were not particularly waiting to be done, but were either very low priority but suggested by having the right resources available. Whilst more goal-oriented models of tasks assume one gathers resources to perform a task, in many real-world situations this gathering of resources is the hard or expensive thing, and it is worth doing activities using the resources and location available, even preparatory ‘just in case’ activities.
DESIGN CONSIDERATIONS

It is worth noting that the user of a ubiquitous device often has to focus on more than one task because s/he might have to interact with the device itself (which is itself a task) while probably performing another task in the real world setting (where this could be the primary task or the secondary task). On one hand, interaction with the ubiquitous device/application to some extent requires user’s innate resources (such as attention). On the other hand, the latter task often too does require the user’s physical, visual, and cognitive involvement/resources (such as hands, visual attention, mental focus). The user’s physical, visual, and cognitive involvement/resources are therefore likely to get constrained. Ideally, the ubiquitous application (including interactions with the device) should support the user in carrying out that which is the primary task without ‘supporting’ the user in tampering with the primary task. We should minimize distracting the user from the primary task or disrupting the user’s primary task, unless the disruption/distraction is of genuine (and great) value or of critical importance. In the words of Holland and Morse: “It is important that the critical focus of the user’s attention be directed towards the primary task at hand” (Holland & Morse, 2001). In adopting ways to meet the requirement, it is also critical to consider the status of a user’s attention in the timing of the tasks on the ubiquitous device. Borrowing from a research effort on guidelines for using agents and direct manipulation (Horvitz, 1999), it is important to “consider the costs and benefits of deferring action to a time when action will be less distracting.” Where necessary, the ubiquitous application should enable/allow the user to temporarily halt a task on the device and to resume the interrupted task.

One of the challenges with a new or innovative technology/application is that its users may try to use it in situations or ways the designers and developers had never thought of. This is true in mobile computing (Gorlenko & Merrick, 2003). There is therefore a sense in which the user may perform tasks on the device (and otherwise) in unpredictable and opportunistic ways. Taking into account all possible scenarios of use for a product is a non-trivial challenge to the ubiquitous application analysts and designers. It is also worth observing that the variability of the environment/natural setting may affect the course of a task. Therefore, analysts and designers may also need to account for such variability in the task analysis (Gorlenko et al., 2003).

The model human processor model (Card et al., 1983) has been a benchmark for a lot of work in HCI. The model is a simplified view of the human processing while interacting with computers. It focuses on the internal cognition driven by the cooperation of the perceptual system, the motor system, and the cognitive system. Each of the systems maintains its own processing and memory. However, as the role and domain of the computers (and devices) have widened, researchers and designers have been considering theories and approaches that take into account the relationship between the internal cognition and the outside world (Dix et al., 2004). Among these, researchers are exploring the following three main understandings of cognition for possible application in ubiquitous computing; they are activity theory model, situated action model, distributed cognition model (Abowd et al., 2002), and even their variants.

Activity Theory

The activity theory model provides a broad conceptual framework for describing the structure, development, and context of computer-supported activities. It was developed by the Russian psychologists Vygotsky, Rubinshtein, Leont’ev and others (Kaptelinin et al., 1995; Leont’ev, 1978). Activity theory is comprised of a set of basic
principles that constitute a general conceptual system, rather than a highly predictive theory. The principles include the hierarchical structure of activity, object-orientedness, internalization/externalization, tool mediation, and development. It should be noted that the principles should be considered as an integrated system, because they are associated with various aspects of the whole activity. In activity theory, the unit of analysis is an activity. The activity is directed at an object which motivates the activity, giving it a specific direction. An activity is made up of goal-directed actions that must be undertaken to fulfill the object. Different actions may be undertaken to meet the same goal. Actions are conscious and they are implemented through automatic operations. Operations do not have their own goals, but rather they provide an adjustment of actions to suit current situations. Therefore, the constituents of activity are not fixed, but can dynamically change (or adjust) as situations/conditions change. This principle is of great interest in ubiquitous computing, since it is desired that the ubiquitous application appropriately adapt to the changing conditions/context. In the context of activity theory, the principle of object-orientedness states that human beings live in a reality that is objective in a broad sense; the things that constitute this reality have not only the properties that are considered objective according to natural sciences, but also socially/culturally defined properties as well. The principle of object-orientedness is very relevant to ubiquitous computing since the ubicomp to a great extent leads to situations where the user directly interacts with other people while (at the same time) using the ubicomp device or application. Therefore the social and cultural issues become even more crucial. An example is society’s perspective regarding a person’s speaking on a cellphone while directly interacting with another person. Internalization is the transformation of external activities into internal ones. Externalization transforms internal activities into external ones. Activity theory emphasizes that internal activities cannot be understood if they are analyzed separately from external activities, because they transform into each other. The external activities in this case can be closely associated with the contextual aspects in ubiquitous computing. For instance, the way the activity of speaking on the cellphone is designed could be better informed by considering the contextual aspects such as the simultaneous but direct interaction with another person, the noise level in the locality, and so forth. Activity theory emphasizes that human activity is generally mediated by tools. Tools are created and transformed during the development of the activity itself. Tools carry with them a particular culture, and therefore the use of tools is an accumulation and transmission of social knowledge. In ubiquitous computing, such tools could in a way be viewed as the ubicomp devices and applications. As far as activity theory is concerned, development is not only an object of study, but also a general research methodology. Gay and Hembrooke have noted a weakness in the original formulation of the activity theory model by pointing out that the model “has traditionally been understood as asynchronic, point-in-time depiction of an activity” (Gay & Hembrooke, 2003).

They go on to note that the model “does not depict the transformational and developmental processes that provide the focus of much recent activity theory research” (Gay & Hembrooke, 2003). In (Boer et al., 2002), Boer et al. do propose an extension of activity theory across time and the levels of an organization to explain connections between different activities as well as the influence that an activity may have on itself. Moreover, Boer et al. also consider the role that an activity may play in other activities at different levels of analysis. Those extensions to the activity theory can serve at least two purposes; they can help to explain tensions present in real-world systems and yield a model with a greater degree of agility in representing complex, distributed cognition. Other work (Uden, 2007) describes how activity theory was used to develop a framework for the
Designing for Tasks in Ubiquitous Computing

Activity theory has also been used in the arena of peripheral displays.

In (Matthews et al., 2007), activity theory was used to perform an analysis of peripheral displays. In the same effort, the authors also used activity theory to develop an approach for designing and evaluating peripheral displays.

Situated Action

The situated action model emphasizes the emergent, contingent nature of human activity, that is, the way activity grows directly out of the particularities of a given situation. The focus is situated activity or practice. The situated action model does not underestimate the importance of artifacts or social relations or knowledge or values, but rather its true locus of inquiry is the “everyday activity of persons acting in [a] setting” (Lave, 1988). The world of computing has always faced contextual issues. However, the current wide adoption and usage of ubiquitous computing (e.g., cellphones, personal digital assistants, etc.) have made contextual issues arguably more prominent than during any other time in history of computing. The main reason is that the ubiquitous devices and applications primarily are used in real settings and therefore, there is a need for the ubiquitous devices and applications to support situated activities. The basic unit of analysis in situated action models is “the activity of persons-acting in setting.” (Lave, 1988). The unit of analysis is thus neither the individual, nor the environment, but rather a relation between the two. The situated action model stresses responsiveness to the environment and the improvisatory nature of human activity. Users under the influence of the environment, may use or attempt to use ubiquitous technologies/applications in “new” ways that even the designers had not anticipated. The situated action model, therefore, can be suitable for capturing and accommodating such user improvisations. On the same note, the situated
action model deemphasizes the study of more durable, stable phenomena that persist across situations (Nardi, 1996). A central tenet of the situated action approach is that the structuring of activity is not something that precedes it, but can only grow directly out of the immediacy of the situation (Nardi, 1996; Lave, 1988). The authors of the effort (Fithian et al., 2003) report that they mainly used the situated action model during the design and evaluation of an integrated location-aware event and meeting planner built to work in a PDA form factor. Their justification for adopting the situated action model was that they “wished to examine the behavior and performance of users in real-world situations, where environmental and social factors are a source of both distraction and motivation” (Fithian et al., 2003; Taylor & Harper, 2002). Fithian et al. actually attribute their meaningful evaluation results to their choice of the situated action model.

Distributed Cognition

Flor et al. in (Flor et al., 1991) describe distributed cognition as “a new branch of cognitive science devoted to the study of: the representation of knowledge both inside the heads of individuals and in the world ...; the propagation of knowledge between different individuals and artifacts ...; and the transformations which external structures undergo when operated on by individuals and artifacts.... By studying cognitive phenomena in this fashion it is hoped that an understanding of how intelligence is manifested at the systems level, as opposed to the individual cognitive level, will be obtained.” It should be observed that ubiquitous devices and applications are primarily used within real settings/context (the world). Therefore, it is important that knowledge pertaining to the real settings be modeled. As has been the case with the desktop computing applications, knowledge about the target user too is important in the arena of ubiquitous computing. On the same note, it is worth noting that the users of ubiquitous technolo-

gies tend to operate in real settings and, therefore, often have to simultaneously interact with other people/individuals and artifacts. Knowledge pertaining to such artifacts and such other individuals is, therefore, important to the design and development of the ubiquitous applications and devices being used. In distributed cognition, the unit of analysis is a cognitive system composed of individuals and the artifacts they use (Flor et al., 1991). Distributed cognition moves the unit of analysis to the system and finds its center of gravity in the functioning of the system (Nardi, 1996). In a manner similar to traditional cognitive science (Newell et al., 1972), distributed cognition is concerned with structure (representations inside and outside the head) and the transformations these structures undergo. However, the difference is that cooperating people and artifacts are the focus of interest, not just individual cognition “in the head” (Nardi, 1996). Another aspect that distributed cognition emphasizes is the understanding of the coordination among individuals and artifacts. The work reported in (Spinelli et al., 2002) is an investigation of users involved in carrying out collaborative activities, locally distributed and mobile. The investigation utilizes the distributed cognition framework and contextual design for representing and analyzing the work observed. By using distributed cognition to model cognition across users and artifacts, the study could look at collaboration from an innovative point of view that highlights how context and external resources impact collaboration. In (Laru & Järvelä, 2003), the authors address an effort that has used distributed cognition and collaborative learning in order to develop a pedagogical model of mobile learning. UbiLearn is a ubiquitous and mobile learning project (Laroussi, 2004). Its work is based on two mobile learning viewpoints; the first is the technical oriented perspective which focuses on a traditional behaviouristic educational paradigm as given and tries to represent or to support it with mobile technologies. The second is the pedagogical socio-cognitive and
distributed cognition paradigms, where we face traditional designs of teaching and learning to push community oriented learning (e.g., collaborative learning, problem based learning; informal and ad-hoc learning, etc.). The work (Fischer et al., 2004) explores the concept of distributed cognition in ubiquitous computing from two directions. On the one hand, it explores the unique possibilities that computational media can have on distributed cognition (how ubicomp technologies can be used to support the users’ distributed cognition). On the other hand, it describes a set of interrelated socio-technical developments that support distributed cognition among communities in ubicomp environments, such as a mobile architecture that links mobile travelers with caregiver communities and transportation systems. The architecture embodies “a distributed cognition framework that avoids common cognitive barriers found in current transportation systems (i.e., generic maps, schedules, labels, landmarks and signs) while synthesizing personalized multimodal attention and memory prompts from the transportation environment to provide travelers with the right information, at the right time, and in a form best suited for the individual traveler” (Fischer et al., 2004).

**Situated Interaction**

It may be resourceful to highlight an interaction paradigm, namely situated interaction that has been defined based on and motivated by some of the above models. Situated interaction refers to the integration of human-computer interaction and the user’s situation in a particular working context in a mobile environment (Hewagamage & Hirakawa, 2000). This combination perceives that the interaction is not only a function of device, but also strongly dependent on the user’s activities and context in which the device is used. The concept of situated interaction can be discerned in, and may be said to have been inspired by, both the situation action model and the activity theory model. Situated interaction actually introduces a new paradigm of computing by extending the conventional applications and also by creating a new set of applications. It is worth noting that mobile computing has become popular in enhancing the shopping experience as discussed in (Newcomb et al., 2003) where they utilized ideas from situated computing. They go on to say that understanding situated interactions, where the customer utilizes the user interface while shopping, became the greatest challenge for designing the ubiquitous user interface.

It is worth noting that the acknowledgement of such challenges could also be supported by the adoption of design approaches and methodologies inspired by participatory design, which is based on the observation of users’ activities in authentic everyday settings where mobile computing takes place (Rogers et al., 2002; Strömberg et al., 2004); as an example, micro-learning is an emergent area of investigation that could find useful resources in these methods while addressing its objective of designing and distributing series of very small units of knowledge to be experienced by learners (for lifelong learning purposes) as intertwined in their everyday working practices and ubiquitous computing activities (Gabrielli et al., 2005).

**EVALUATION CONSIDERATIONS**

Conventional user-centered methods could be appropriately exploited in the development process of ubiquitous applications. On the same note, some of the traditional usability evaluation techniques might become useful when adapted for ubiquitous computing. For instance, there are several efforts toward realizing usability principles and heuristics for the design and evaluation of ubiquitous environments/systems, such as ambient heuristics (Mankoff et al., 2003) and groupware heuristics (Baker et al., 2001). On the same note, we actually already have proposed a review of usability principles for mobile computing (Bertini et al.,
Designing for Tasks in Ubiquitous Computing

2005). We have also developed usability heuristics that are appropriate for evaluation in mobile computing (Bertini et al., 2006).

Much traditional understanding of work organizations has its roots in Fordist and Taylorist models of human activity, which assume that human behavior can be reduced into structured tasks. HCI has not been spared from this either. In particular, evaluation methods in HCI have often relied on measures of task performance and task efficiency as a means of evaluating the underlying application. However, it is not clear whether such measures can be universally applicable when we consider the current move from rather structured tasks (such as desktop activities) and relatively stable settings to the often unpredictable ubiquitous settings. Such primarily task-centric evaluation may, therefore, not be directly applicable to the ubiquitous computing domain. It would be interesting to consider investigating methods that go beyond the traditional task-centric approaches (Abowd & Mynatt, 2000). It is also worth keeping in mind that tasks on the ubiquitous device (and elsewhere) tend to be unpredictable and opportunistic.

In this era of ubiquitous computing, the real need to take into account the real-world context has become more crucial than at any other time in the history of computing. Although the concept of context is not new to the field of usability (e.g., ISO 9241 guidelines propose a “model” consideration of context), evaluation methods have, however, found it challenging, in practice to adequately/completely integrate the entire context during the evaluation process. There are various ways to address this challenge.

One option is the employment of observational techniques (originally developed by different disciplines) to gain a richer understanding of context (Abowd et al., 2002; Dix et al., 2004). Main candidates are ethnography, cultural probes, and contextual design. Another option is to use the “Wizard-of-Oz” technique, other simulation techniques, or even techniques that support the participant’s imagination. Prototyping too presents an avenue for evaluating ubiquitous computing applications.

Ethnography

Ethnography is an observational technique that uses a naturalistic perspective; that is, it seeks to understand settings as they naturally occur, rather than in artificial or experimental conditions, from the point of view of the people who inhabit those settings, and usually involves quite lengthy periods of time at the study site (Hughes et al., 1995). Ethnography involves immersing an individual researcher or research team in the everyday activities of an organization or society, usually for a prolonged period of time. Ethnography is a well-established technique in sociology and anthropology. The principle virtue of ethnography is its ability to make visible the ‘real world’ aspects of a social setting. It is a naturalistic method relying upon material drawn from the first-hand experience of a fieldworker in some setting. Since ubiquitous devices and applications are mainly used in ‘real world’ settings, then ethnography has some relevance to ubiquitous computing. The aim of ethnography is to see activities as social actions embedded within a socially organized domain and accomplished in and through the day-to-day activities of participants (Hughes et al., 1995). Data collected/gathered from an ethnographic study allows developers to design systems that take into account the sociality of interactions that occur in the “real world.” The work by Crabtree et al. (Crabtree et al., 2006), shows how ethnography is relevant to and can be applied in the design of ubiquitous computing applications. The ultimate aim of the effort is to “foster a program of research and development that incorporates ethnography into ubiquitous computing by design, exploiting the inherent features of ubiquitous computing applications to complement existing techniques of observation, data production, and analysis.” While describing
how mobile computing has been used in the fashion retail industry, Supawanich et al. highlight challenges such as those pertaining to usability, system tailoring, and the manager-client user experience (Supawanich et al., 2005). It is worth noting that they applied ethnography toward addressing the foregoing challenges. In the work (Newcomb et al., 2003), which we have mentioned before, the authors also have applied ethnography in their effort to examine how grocery shopping could be aided by a mobile shopping application for the consumers. In particular, the authors shopped with customers and followed them throughout the task of shopping, observing their shopping habits. In (Berry & Hamilton, 2006), the authors report that they used ethnography in order to understand multimedia students and how they use Tablet PCs in their everyday design studies.

**Cultural Probes**

Cultural probes (Gaver et al., 1999a) represent a design-led approach to understanding users that stresses empathy and engagement. They were initially deployed in the Presence Project (Gaver et al., 1999b), which was dedicated to exploring the design space for the elderly. Gaver has subsequently argued that in moving out into everyday life more generally, design needs to move away from such concepts as production and efficiency and instead focus and develop support for “ludic pursuits.” This concept is intended to draw attention to the “playful” character of human life, which might best be understood in a post-modern sense. Accordingly, the notion of “playfulness” is not restricted to whatever passes as entertainment, but is far more subtle and comprehensive, directing attention to the highly personal and diverse ways in which people “explore, wonder, love, worship, and waste time” together and in other ways engage in activities that are “meaningful and valuable” to them (Gaver, 2001). This emphasis on the ludic derives from the conceptual arts, particularly the influence of Situationist and Surrealist schools of thought (Gaver et al., 1999a). Cultural probes draw on the conceptual arts to provoke or call forth the ludic and so illuminate the “local culture” in which people are located and play out their lives. During their course of use, ubiquitous devices and applications typically get embedded in the users’ lives and cultures. For instance, people often get personally attached to their cellphones. Cultural probes offer fragmentary glimpses into the rich texture of people’s lives (Gaver, 2002). Cultural probes are not analytic devices but “reflect” the local culture of participants and are drawn upon to inspire design. In the Presence Project, cultural probes inspire design by providing a rich and varied set of materials that help to ground designs in the detailed textures of the local cultures (Gaver et al., 1999a). These materials are products of the probe packs, each consisting of a variety of artifacts relevant to the study. Such artifacts provide a range of materials reflecting important aspects of the participant’s local cultures and, on being returned to the investigators, these reflections inspire design. For instance, in the Presence Project, the artifacts include: postcards with questions concerning participants’ attitudes to their lives, cultural environment, and technology; maps asking participants to highlight important areas in their cultural environment; cameras with instructions asking participants to photograph things of interest to them and things that bored them; photo albums asking participants to assemble a small montage telling a story about participant’s lives; and media diaries asking participants to record the various media they use, when, where, and in whose company. The original idea of culture probes has been extended to include technology and thus the concept, technology probes (Hutchinson et al., 2003; Paulos & Goodman, 2004; Paulos & Jenkins, 2005). According to (Hutchinson et al., 2003), technology probes can assist in achieving “three interdisciplinary goals: the social science goal of understanding the needs and desires of users in a real-world setting; the engineering goal of field testing the technology; and the design goal of in-
Designing for Tasks in Ubiquitous Computing

spiring users and researchers to think about new technologies.” It is also possible to consider a probe that is entirely simulated, such as with paratypes (Abowd et al., 2005). In a research effort aimed at exploring issues of dependability in ubiquitous computing in domestic settings (Crabtree et al., 2002), cultural probes are one of the qualitative methods that was used. In this case, some participants agreed to keep personal diaries of their daily activities. However, all participants were supplied with polaroid cameras, voice activated dictaphones, disposable cameras, photo albums, visitors books, scrapbooks, post-it notes, pens, pencils and crayons, postcards, and maps. In an attempt to elicit the methods and guidelines for designing and developing applications for domestic ubiquitous computing, Schmidt and Terrenghi in (Schmidt & Terrenghi, 2007) adopted various methods including cultural probes. In a study of the possible applications of mobile technology for industrial designers and architects for their daily work, Muñoz Bravo et al. in (Muñoz Bravo et al., 2007) conducted user studies in which one of the studies consisted of using cultural probes.

Contextual Inquiry

Contextual inquiry (Holtzblatt et al., 1993) is a method that aims at grounding design in the context of the work being performed. Contextual inquiry recommends the observation of work as it occurs in its authentic setting, and the usage of a graphical modeling language to describe the work process and to discover places where technology could overcome an observed difficulty. It is worth noting that in its application, contextual inquiry does combine various methods such as field research and participatory design methods (Muller et al., 1993) in order to provide designers with grounded and rich/detailed knowledge of user work. Contextual inquiry is one of the parts of what is referred to as contextual design. Contextual design is a design approach that was developed by Holtzblatt and Beyer (Beyer et al., 1998). It is an approach for designing customer-centered products based on an understanding of the existing work contexts and practices. It is worth noting that ubiquitous devices and applications are often intended to be used and get used in the real world where real work (or primary tasks) take(s) place. Therefore, the design of such devices and applications should be informed by an understanding of the way customers work (or would like to work) in the real world. Contextual design starts with the premise that any product embodies a way of working. The product’s function and structure introduce particular strategies, language, and work flow on its users. A successful design should therefore offer a way of working that customers would like to adopt. Contextual design has seven parts: contextual inquiry; work modeling; consolidation; work redesign; user environment design; testing with customers; and putting it into practice. One of the proponents of contextual design, Holtzblatt, has actually reported on how contextual design can be appropriated to produce a mobile application (Holtzblatt, 2005). It is interesting to observe that the work by Newcomb et al. (Newcomb et al., 2003), did come up with a contextual design which was meant to serve two purposes; these are in the shopper’s home to aid him/her in creating a shopping list, and in the store for the actual shopping. In the effort by Schmidt and Terrenghi (Schmidt & Terrenghi, 2007), which we came across before, contextual inquiry too was used for understanding and proposing methods and guidelines for designing and developing domestic ubiquitous computing applications. The previously mentioned work by Spinelli et al. (2002) on locally distributed and mobile collaborative activities, which we came across before, did use contextual design. The authors defend their choice of contextual design by stating that “the representation of work activities, utilising the methods of contextual design, aid researchers in conceptualising technologies that truly meet the informational and communicative needs of dynamic and fragmented users
and their communities. ... This has allowed us to develop an understanding of, and to design for, users and their communities-in-context, by applying techniques such as affinity diagramming (for theme building) and work models to capture such essential elements as cultural and social models of technology use; ‘breakdowns’ ... in working practices and artefact models ... that allows us to represent users resources and their relationship with these resources. In the process, it also promotes an effective coupling of well-designed technologies with the fast changing physical environments that their users may inhabit” (Spinelli et al., 2002).

‘Wizard-of-Oz’ Simulation and Supporting Immersion

Another possibility is to use the “Wizard-of-Oz” technique or even other simulation techniques such as virtual reality. The “Wizard-of-Oz” technique is an evaluation method where the user of the system is made to believe or perceive that he or she is interacting with a fully implemented system though the whole or a part of the interaction of the system is controlled by a human being, the “wizard,” or several of them. Such techniques are especially appropriate where the ubiquitous application is not fully complete. However, the simulation should closely reflect the real context as much as possible (realistic simulation). There exist various ubiquitous computing applications that have at some point been evaluated using the “Wizard-of-Oz” technique, for example, (Carter et al., 2007; Mäkelä et al., 2001; Rudström et al., 2003), and so on. Another alternative is to adapt more traditional inspection methods to the analysis of ubicomp settings by enriching the range and quality of discovery resources provided to usability experts to support their imagination and immersion about the real world usage settings. We have recently conducted a study in this direction where video data about user interaction with an e-learning course delivered on PDAs were used as additional resources supporting a more effective performance of cognitive walkthrough evaluation by usability experts involved in the study (Gabrielli et al., 2005).

Prototypes

In the formative stages of the design process, low fidelity prototypes can be used. However, as the design progresses, user tests need to be introduced. In the context of ubiquitous computing, user tests will not only require the inclusion of real users, real settings, and device interaction tasks, but also real or primary tasks (or realistic simulations of the real tasks and of the real settings). As mentioned previously, realistic simulations of the real tasks and of the real settings could be adopted as an alternative. Therefore, there would be the need to provide a prototype that supports the real tasks and real settings or their simulations. This does imply some cost in the design process because the prototype at this level would need to be robust and reliable enough in order to support primary tasks in real settings or the simulations. In fact, the technology required to develop ubiquitous computing systems is often on the cutting edge. Finding people with corresponding skills is difficult. As a result, developing a reliable and robust ubiquitous computing prototype or application is not easy (Abowd & Mynatt, 2000; Abowd et al., 2002).

OPEN ISSUES AND CONCLUSION

We have attempted to describe the nature of tasks in ubiquitous computing. We have then proposed and discussed various models and methods appropriate for supporting the development process of ubiquitous computing applications based on the deeper understanding of the nature of tasks. However, still there are many other pertinent aspects which too would need to be addressed and which we consider worthy of our further
Investigation. These include: the choice of the methods; the choice of the models; the classification/categorization and characterization of tasks for mobile and ubiquitous computing; formal specification of social and collaborative aspects; and so forth.

**Choice of Methods**

We have described several methods appropriate for evaluating in ubiquitous computing. One of the major issues is deciding which of the methods to choose. Of such evaluation methods, one may want to know which one(s) will be most suitable for a certain ubicomp application. Considering evaluation methods in general (not just evaluation methods for ubicomp), Dix et al. indicate that: “there are no hard and fast rules in this – each method has its particular strengths and weakness and each is useful if applied appropriately.” (Dix et al., 2004). They, however, point out that there are various factors worth taking into consideration when choosing evaluation method(s), namely:

- the stage in the lifecycle at which the evaluation is carried out;
- the style of evaluation (field or laboratory);
- the level of subjectivity or objectivity of the method;
- the type of measures provided by the method;
- the level of information provided by the method;
- the immediacy of the response provided by the method;
- the level of interference or intrusiveness of the method;
- the resources required by the method.

The foregoing factors may be appropriately borrowed from when we consider the evaluation of ubicomp applications. According to Carter et al. (Carter et al., 2007), in determining which methods for ubiquitous computing to use, (among paper prototypes, interactive prototypes, “Wizard-of-Oz,” and probes,) the designer must make trade-offs between realism, unobtrusiveness, data sparsity, ambiguity, and cost/time. They go on to say that paper prototypes and “Wizard-of-Oz” can be used to explore ambiguity. Probes that can be employed in real-world situations over a period of time can support both realism and sparsity. Moreover, paper and interactive prototypes may be the least costly methods, but they may also be the least flexible methods. It therefore comes as no surprise that some researchers have begun carrying out corresponding comparative studies (Liu & Khooshabeh, 2003; Mankoff & Schilit, 1997).

Many of the methods considered in this chapter are very “open” compared to more traditional task analysis techniques. This reflects the often spontaneously planned and re-planned nature of many tasks “in the wild” compared to (relatively) more constrained office tasks. Methods that embody a fixed or pre-understood idea of human behaviour are likely to miss some of the nuanced activity that is the focus of more open observational techniques such as ethnography. However, without models it is hard to move from what is observed to potential, especially as this potential often involves users appropriating technology for themselves. For this prompting to see what could happen, as well as what does happen, more interventionist methods in particular forms of technology probes or at least rich prototypes seem more appropriate. That is, the more open methods seem best suited for early and late stages in design for understanding the initial situation and later for assessing the impact of a deployment. However in mid-stages, when establishing potential is more important, more structured models and more interventionist methods seem more appropriate.
Choice of Models

Fithian et al. in (Fithian et al., 2003) observe that mobile and ubiquitous computing applications lend themselves well to the models: situated action; activity theory; and distributed cognition. As for which of these models are most suitable for a certain mobile or ubiquitous application, the foregoing authors say that the choice depends largely on the kind of application and of which aspects of design are in the limelight. They recommend that the choice be based on a critical analysis of the users and their knowledge, the tasks, and the application domain.

In (Fithian et al., 2003), Fithian et al. also note that basing entire evaluation on just time measurements can be very limiting, especially if the tasks are benchmarked in a situated action setting. Although time measurements are important, other performance measures that may be much more useful for evaluating such ubicomp applications include interruption resiliency, interaction suspensions, interaction resumptions, and so forth.

Interestingly, these richer metrics require a far richer model of what is going on than simpler end-to-end timing. This reinforces the message on other areas of evaluation that understanding mechanism is critical for appropriate and reliable generalization (Ellis & Dix, 2006).

Classification of Tasks

In a study found in (Carter et al., 2007), Carter et al. report that respondents felt that the current mobile tools are poorly matched to the user tasks of meeting and “keeping up with” friends and acquaintances. The study observed that location-based technology might assist users in such tasks. Moreover, the study found that users would prefer to have cumbersome and repetitive tasks carried out by their mobile technology artifacts (e.g., the device, the application, etc.). Carter et al. also found that planning tasks vary in nature and detail depending on the formal or informal nature of the event. It might be interesting to consider how level of formality could be used as one of the means of classifying tasks in mobile and ubiquitous computing. Carter et al. observe that events with differing levels of formality require different tasks and, therefore, different support. They note that users showed most interest in the systems that supported informal gathering, rather than formal gatherings. Another possible criterion for classifying or categorizing user tasks could be by borrowing from the activity theory’s framework for describing human behavior (e.g., activities, operations, actions, etc.) or more specialized frameworks such as (Bardram, 2005; Bardram & Christensen, 2004). The work (Matthews et al., 2007), proposes the following classification for the types of activities peripheral displays are likely to support: dormant, primary, secondary, and pending.

Characterization of Tasks

In a work which primarily describes the challenges for representing and supporting user’s activity in the desktop and ubiquitous interactions, Voida et al. in (Voida et al., to appear) characterize activities as follows:

- activities are dynamic, emphasizing the continuation and evolution of work artifacts in contrast to closure and archiving;
- activities are collaborative, in the creation, communication, and dissemination of work artifacts;
- activities exist at different levels of granularity, due to varying durations, complexity and ownership; and
- activities exist across places, including physical boundaries, virtual boundaries of information security and access, and fixed and mobile settings.

In (Abowd & Mynatt, 2000), Abowd and Mynatt describe everyday computing as an area
Designing for Tasks in Ubiquitous Computing

of interaction research which results from considering the consequences of scaling ubiquitous computing with respect to time. They indicate that designing for everyday computing requires focus on the following features of informal, daily activities:

- they rarely have a clear beginning or end;
- interruption is expected;
- multiple activities operate concurrently;
- time is an important discriminator;
- associative models of information are needed.

Like Fithian et al.’s metrics described above (Fithian et al., 2003), these properties all emphasize the fact that activities in a ubiquitous interaction are more fragmented and require more divided attention than “architypal” office applications, although arguably these were never as simple as the more simplistic models suggested. However, the first point also suggests that at a high-level there may be more continuity, and this certainly echoes Carter et al.’s study (Carter et al., 2007) with the importance of informal gathering and communication a life-long goal.

Formal Specification of Social and Collaborative Aspects

With a formal specification, it is possible to “analyze” a system long before it gets designed or implemented. Although this benefit applies to virtually all types of systems, it is interesting to the world of ubiquitous computing where, as we have noted, at the end of the previous section, developing a reliable and robust prototype or application is not an easy undertaking. Formal specifications, therefore, can be useful in supporting the development of ubiquitous applications. On the same note, in ubiquitous computing users perform their activities in the real world settings, where there are other people. In other words, ubiquitous computing involves context, which includes other people besides the user. Therefore, collaborative and social aspects have a lot of weight in ubiquitous computing. It has been rightly noted in (Abowd & Mynatt, 2000) that human beings tailor their activities and recall events from the past based on the presence (or even the help) of other people. Therefore, it is important to consider how we can realize formal specifications that can represent collaborative and social aspects for ubiquitous applications. It is worth observing that much of the research in ubiquitous computing has focused on mobility (and other contextual aspects) with regard to an individual user, with little being done regarding social and collaborative aspects.

One of the possible approaches to the formal modeling of social aspects is through the use of agents. It might be worth investigating to what degree such agent-based models can be applied in ubiquitous computing. One such model is OperA (Dignum et al., 2002a; Dignum et al., 2002b; Dignum, 2004). The authors indicate that the concept of agents is useful for representing organizational interaction for two main reasons. The first is that it enables the reference to any autonomous entity participating in an interaction, including people. The second is that it provides theoretical models for entities and interaction. OperA “abstracts from the specific internal representations of the individual agents, and separates the modeling of organizational requirements and aims. Contracts are used to link the different models and create specific instances that reflect the needs and structure of the current environment and participants” (Dignum, 2004).

It might also be appropriate to borrow a leaf from Grid Computing where the need for models for addressing collaborative and social aspects has been identified (Liu & Harrison, 2002). According to Liu (Liu, 2003), in Grid Computing the development of such models has been based on the early work on information systems (Liu, 2000; Stamper, 1973; Stamper, 1996) and computer-supported collaborative work (CSCW),
Designing for Tasks in Ubiquitous Computing

(Liu et al., 2001). One particular model that has been proposed is the SPS model, which entails the integrated modeling of semantic, pragmatic and social aspects (Liu, 2003). Regarding formal specifications for CSCW in general (and not just under Grid Computing), one interesting effort is the work by Johnson (Johnson, 1999), which describes how formal methods can be used creatively to solve a vast range of design problems within CSCW interfaces. It is worth noting that the work does show how mathematical specification techniques can be enhanced to capture physical properties of working environments, thereby providing a link between the physiological studies from ergonomics and the HCI user interface design techniques.

A related and interesting work is found in (Musolesi et al., 2004), in which there is a proposal of a two-level mobility model that is based on artificially generated social relationships among individuals carrying mobile devices. The generation process respects the mathematical basis of social networks theory and, thus, is grounded in empirical experience of actual social relationships. The second level/stage maps the social organization onto topographical space such that the actual generated topography is biased by the strength of social ties.

At a very low level, more traditional formal models become applicable as we are “below” the level of the more complex considerations of ubiquitous computing. In particular, variations of Fitts’ law have been used extensively to understand and to design interfaces for pointing tasks on tiny devices (Guiard & Beaudouin-Lafon, 2004).

**SUMMARY**

As a way of emphasizing the relevance of the theme of this chapter, it is worth observing that there is a growing interest within the research community regarding tasks in ubiquitous computing. Therefore, it comes as no surprise that we are now seeing the emergence of fields such as activity-centered design (Gay & Hembrooke, 2003), activity-based computing (Bardram, 2005; Bardram & Christensen, 2004), and activity-based ubiquitous computing (Li & Landay, 2006).

As we consider the move from the conventional desktop setting to the real world setting, various design issues and demands arise when we consider the nature of tasks the ubiquitous devices/applications would be expected to support and the real world context in which they will be used. A close study of the nature of tasks in ubiquitous computing has the potential to bring to light some of the requirements in the development of ubiquitous applications.

In particular, we have seen how tasks in ubiquitous environments tend to be more dynamic, less pre-planned, and more situated than those commonly assumed to be the case for more traditional desktop applications. In addition, users are likely to be involved in multiple activities, and the task involving a ubiquitous device may not be the primary task for the user either because there is a real world task(s) such as driving that takes precedence, or because the device interaction is merely supporting an ongoing activity such as social coordination. Interruptions and resumptions of activity become the norm (although there is plenty of evidence that this is also the case in the office) and so the need, as advocated in distributed cognition, to offload memory into the device becomes important.

Because of the dynamic nature of tasks we have discussed, various methods and theories that emphasise the richer nature of human activity, and any methods used to study tasks for ubiquitous interaction have to be open to seeing unexpected patterns of activity. However, there are clearly also generic meta-tasks and common issues found in many ubiquitous interactions including offloading of memory, interruption management, location sensitivity, and so forth. It is essential to understand the former, situation specific issues, in order to avoid designs that
are not fit for that purpose; however, the latter, generic issues, offer the potential for lessons to be learnt across systems and for ongoing fruitful research directions.

REFERENCES


*This work was previously published in Advances in Ubiquitous Computing: Future Paradigms and Directions, edited by S. Mostefaoui, Z. Maamar, and G. Giaglis, pp. 171-200, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
INTRODUCTION

In ontological engineering research field, the concept of “task ontology” is well-known as a useful technology to systemize and accumulate the knowledge to perform problem-solving tasks (e.g., diagnosis, design, scheduling, and so on). A task ontology refers to a system of a vocabulary/concepts used as building blocks to perform a problem-solving task in a machine readable manner, so that the system and humans can collaboratively solve a problem based on it.

The concept of task ontology was proposed by Mizoguchi (Mizoguchi, Tijerino, & Ikeda, 1992, 1995) and its validity is substantiated by development of many practical knowledge-based systems (Hori & Yoshida, 1998; Ikeda, Seta, & Mizoguchi, 1997; Izumi &Yamaguchi, 2002; Schreiber et al., 2000; Seta, Ikeda, Kakusho, & Mizoguchi, 1997). He stated:

...task ontology characterizes the computational architecture of a knowledge-based system which performs a task. The idea of task ontology which serves as a system of the vocabulary/concepts used as building blocks for knowledge-based systems might provide an effective methodology and vocabulary for both analyzing and synthesizing knowledge-based systems. It is useful for describing inherent problem-solving structure of the existing tasks domain-independently. It is obtained by analyzing task structures of real world problem.

... The ultimate goal of task ontology research is to provide a theory of all the vocabulary/concepts necessary for building a model of human problem solving processes. (Mizoguchi, 2003)

We can also recognize task ontology as a static user model (Seta et al., 1997), which captures the meaning of problem-solving processes, that is, the input/output relation of each activity in a
problem-solving task and its effects on the real world as well as on the humans’ mind.

BACKGROUND

Necessity of Building Task Ontologies as a Basis of HCI

It is extremely difficult to develop an automatic problem-solving system that can cope with a variety of problems. The main reason is that the knowledge for solving a problem varies considerably depending on the nature of the problems. This engenders a fact that is sometimes ignored: Users have more knowledge than computers. From this point of view, the importance of a user-centric system (DeBells, 1995) is now widely recognized by many researchers. Such framework follows a collaborative, problem-solving-based approach between human and computer by establishing harmonious interaction between human and computer.

Many researchers implement such a framework with a human-friendly interface using multimedia network technologies. Needless to say, it is important not only to apply the design principles of the human interface but also principle knowledge for exchanging meaningful information between humans and computers.

Systems have been developed to employ research results of the cognitive science field in order to design usable interfaces that are acceptable to humans. However, regarding the content-oriented view, it is required that the system can understand the meaning of human’s cognitive activities in order to capture a human’s mind.

We, therefore, need to define a cognitive model, that is, to define the cognitive activities humans perform in a problem-solving/decision-making process and the information they infer, and then systemize them as task ontologies in a machine understandable manner in order to develop an effective human-computer interaction.

Problem-Solving Oriented Learning

A task with complicated decision making is referred to as “Problem-Solving Oriented Learning (PSOL) task” (Seta, Tachibana, Umano, & Ikeda, 2003; Seta & Umano, 2002). Specifically, this refers to a task that does not only require learning to build up sufficient understanding for planning and performing problem-solving processes but also to gain the ability/skill of making efficient problem-solving decisions based on sophisticated strategies.

Consider for example, a learner who is not very familiar with Java and XML programming and tries to develop an XML-based document retrieval system. A novice learner in a problem-solving domain tries to gather information from Web resources, investigates and builds up his/her own understanding of the target area, and makes plans to solve the problem at hand and then perform problem-solving and learning processes. Needless to say, a complete plan cannot be made at once, but is detailed gradually by iterating, spirally, those processes while applying a “trial and error” approach. Thus, it is important for a learner to control his/her own cognitive activities.

Facilitating Learners’ Meta Cognition through HCI

In general, most learners in PSOL tend to work in an ad hoc manner without explicit awareness of meaning, goals and roles of their activities. Therefore, it is important to prompt construction of a rational spiral towards making and performing efficient problem-solving processes by giving significant direction using HCI.

Many researchers in the cognitive science field proposed a concept whereby metacognition plays an important role to acquire and transfer expertise (Brown, Bransford, Ferrara, & Campione, 1983; Flavell, 1976; Okamoto, 1999). Furthermore, repeated interaction loops between metacognition activities and cognition activities play an
important role in forming an efficient plan for problem-solving and learning processes.

Figure 1 shows the plan being gradually detailed and refined along the time axis. Figure 1(a) is a planning process when a learner has explicit awareness of interactions and iterate metacognition activities and cognition activities spirally, while Figure 1(b) is a planning process with implicit awareness of them. In PSOL, monitor and control of problem-solving/learning processes are typical activities of metacognition while their performances are ones of cognition. It is natural that the former case allows efficient plans for problem-solving workflow more rapidly than the latter. Without explicit awareness of interaction loops, a learner tends to get confused and lose his/her way because nested structures of his/her work and new information of the target world impose heavy loads.

Therefore, it is important to implement an HCI framework that enables effective PSOL by positioning a learner at the center of the system as a subject of problem solving or learning, and providing appropriate information to prompt the learner’s metacognition effectively.

**MAIN ISSUES IN TASK ONTOLOGY-BASED HCI**

In this section, we introduce our approach to supporting PSOL to understand task ontology based HCI framework.

Rasmussen’s (1986) cognitive model is adopted as a reference model in the construction of the task ontology for supporting PSOL. It simulates the process of human cognition in problem-solving based on cognitive psychology. Cognitive activity in PSOL is related to this model based on which PSOL task ontology is constructed. This provides a learner with useful information for effective performance of cognitive activity at each state, according to the theoretical framework that was revealed in the cognitive psychology.
Rasmussen’s (1986) Cognitive Model

Figure 2 represents an outline of Rasmussen’s cognitive model known as the ladder model.

Activities in PSOL broadly comprise activities in connection with a problem-solving act and activities in connection with a learning act (see Figure 3 in the next section).

An Activation activity in Rasmussen’s cognitive model corresponds to the situation in which a problem is given in problem-solving activities, or one in which a learner detects change in the real world. An Observe activity corresponds to observing the details of the change or a gap from the problem-solving goal. An Identify activity corresponds to identifying its possible cause. An Interpret activity corresponds to interpreting the influence of the change on problem solving and deciding the problem-solving goal. A Define Task activity corresponds to determine a problem-solving task for implementing it based on the problem-solving goal. A Formulate Procedure activity corresponds to setting up a problem-solving plan to solve the problem-solving task.

Although basically the same correspondence applies in learning activities as the case of problem-solving activities, the object of learning activities, mainly focuses on the state of one’s own knowledge or understanding, that is, metacognition activities. Namely, the Activation activity in Rasmussen’s cognitive model corresponds to detecting the change of one’s own knowledge state. The Observe activity corresponds to observing details or a gap from its own understanding state (goal state) decided as a goal of learning. The Identify activity corresponds to identifying its possible cause. The Interpret activity corresponds to interpreting the influence of its own understanding state, especially the influence on problem solving in PSOL, and deciding the goal of learning. The Define Task activity corresponds to setting up a learning task for implementing it based on the problem-solving goal. The Formulate Procedure activity corresponds to setting up a learning plan to solve the problem-solving task.

Clarifying a correspondence relationship between the cognitive activity by a learner in PSOL and the cognitive activity in Rasmussen’s cognitive model permits construction of a problem-solving-oriented learning task ontology as a basis of human-computer interaction comprehending the properties of PSOL appropriately. Implementing an interaction between a system and a learner based on this allows the system to show effective information to encourage the learner’s appropriate decision-making.

Cognitive Model in Problem-Solving Oriented Learning

Figure 3 shows a cognitive model that captures detailed working processes of a learner. This model is PSOL task specific while Rasmussen’s model is a task independent one. By making the correspondence between these models, we can define an HCI framework based on Rasmussen’s theory.
Figures 3(i) and 3(iii) represent the planning process of the problem-solving plan and learning plan, respectively, and 3(viii) and 3(x) represent problem-solving and learning processes in Figure 1, respectively. Figures 3(v) and 3(vi) represent the monitoring process.

We have presented a problem, say, “developing an XML based document retrieval system” in the upper left corner. Two virtual persons, a problem-solving planner and learning process planner in the learner, play roles of planning, monitoring, and controlling problem-solving, and learning processes, respectively.

With PSOL, a learner first defines a problem-solving goal and refines it to sub-goals which contribute to achieving goal G (Figure 3(i)). They are refined to feasible problem-solving plans (Figure 3(ii)); thereafter, the learner performs them to solve the problem (Figure 3(viii)).

If the learner recognizes a lack of knowledge in the sub goals and performs problem-solving plans, we can generate an adequate learning goal (LG) to get knowledge (Figure 3(iii)) and refine it to learning process plans (Figure 3(iv)). In learning processes (Figure 3(x)), s/he constructs knowledge (Figure 3(iv)) to be required to plan and perform the problem-solving process. Based on constructed knowledge, s/he or he specifies and performs the problem-solving processes (Figure 3(viii)), to change the real world (Figure 3(vii)). The learner assesses gaps among goal states (GS), current goal states (CGS) of problem-solving process plans, and current state (c-state) of the real-world (Figure 3(v)) and ones among learning goal states.
(LGS), current learning goal states (CLGS) of learning process plans and understanding state (Figure 3(vi)). She or he continuously iterates these processes until the c-state of the real world satisfies the GS of problem solving.

It is notable that learners in PSOL have to make and perform not only problem-solving plans, but also learning plans in the process of problem solving. Furthermore, it is important for the learner to monitor real-world changes by performing problem-solving processes and to monitor his/her own understanding states by performing learning processes and checking and analyzing whether states of the real world and understanding states satisfy defined goal states (Figures 3(v) and 3(vi)). The gap between current states and goal states causes the definition of new goals to be dissolved.

Consequently, PSOL impels a learner to perform complicated tasks with heavy cognitive loads. A learner needs to manage and allocate the attentional capacity adequately because of limited human attentional capacity. This explains why a novice learner tends to get confused and lose his/her way.

**Task Ontology for Problem-Solving-Oriented Learning**

Figure 4 presents an outline of the PSOL Task Ontology (Problem-Solving-Oriented Learning Task Ontology). Ovals in the figure express a cognitive activity performed by a learner in which a link represents an “is-a” relationship.

The PSOL task ontology defines eight cognitive processes modeled in Rasmussen’s cognitive model as lower concepts (portions in rectangular box (a) in the figure). They are refined through an is-a hierarchy to cognitive activities on the meta-level (meta activity), and cognitive activities on the object level (base activity). Moreover, they are further refined in detail as their lower concepts: a cognitive activity in connection with learning activities and a cognitive activity in connection with problem-solving activities. Thereby, a conceptual system is constructed that reflects the task structure of PSOL. For example, typical metacognition activities that a learner performs in PSOL, such as “Monitor knowledge state” and “Monitor learning plan,” are systematized as lower concepts of metacognition activities in the Observe activity.

Figure 5 shows a conceptual definition of an act that identifies a possible cause of why a plan is infeasible. All the concepts in Figure 5 have a conceptual definition in a machine readable manner like this, thus, the system can understand what the learner tries to do and what information he/she needs.

Cause identification activities defined include: the actor of the activity is a learner; a learner’s awareness of infeasibility becomes an input (in%symptom in Figure 5); the lower plan of an target plan that the learner tries to make it feasible now is made into a reference information (in%reference in Figure 5). Moreover, this cognitive activity stipulates that a learner’s awareness of causes of infeasibility is output (out%cause in Figure 5). The definition also specifies that the causes of the infeasibility include (axioms in Figure 5): that the sufficiency of that target plan is not confirmed (cause1 in Figure 5); that the feasibility of a lower plan, small grained plan that contributes to realize the target plan, is not confirmed (cause2 in Figure 5); and that the target plan is not specified (cause3 in Figure 5). Based on this machine understandable definition, the system can suggest the candidate causes of infeasibility of the object plan, and the information the learner should focus on.

Making this PSOL task ontology into the basis of a system offers useful information in the situation that encourages appropriate decision-making. This is one of the strong advantages using PSOL task ontology.
Figure 4. A hierarchy of problem-solving-oriented learning task ontology

Figure 5. A definition of “identify c-state of NT in executable plan”
An Application: Planning Navigation as an Example

The screen image of Kassist, a system based on the PSOL Task Ontology, is shown in Figure 6. Kassist is an interactive open learner-modeling environment. The system consists of six panels. A learner describes a problem-solving plan, own knowledge state about the object domain, and a learning process in each panels of (a), (b), and (c), respectively. Furthermore, a learner can describe the correspondence relationship between the problem-solving process placed at (a) and the concept of (b), that is, the correspondence relationship with the knowledge of the object domain required for carrying out the process of (a); and the correspondence relationship between the learning process placed at (c), and the concept of (b), that is, the correspondence relationship with the learning process of (c) which constructs an understanding on the concept of (b). Each shaded node in (b) represents either “knowing” or “not knowing” the concept. A learner can describe the correspondence of the concepts and processes in the object world placed on (a), (b), and (c) with resource (f), used as those reference information, so that appropriate information can be referred to when required.

This provides a learner with an environment in which she or he can externalize and express her own knowledge; it then encourages his or her spontaneous metacognition activities such as the Activation activity and Observation activity in Rasmussen’s model. Moreover, we can implement a more positive navigation function that encourages a learner’s metacognition activity.

Figure 6. Interactive navigation based on problem solving oriented learning task ontology
in the subsequent cognitive process by making ontology the basis of a system.

Consider, for example, this task “Investigate how to develop XML-based document retrieval system”. Assume a situation where a learner does not know how to tackle this task:

i. In this situation, a learner clicks the “Investigate how to develop XML-based document retrieval system” node on (e); among the lower learning plans connected by the “part-of” links, a plan “Learn how to design module structure”, whose feasibility is not secured is highlighted based on the ontology; and a learner is shown the causes of infeasibility with a message “connected lower plan is not feasible”.

ii. Then, it shows the cause as a learner has lack of knowledge to specify the plan.

iii. Moreover, plans influenced by the infeasibility of this learning plan are displayed in the interpretation process.

iv. Here, a problem-solving plan “Design module structure” is highlighted. Such navigation allows a learner to comprehend knowledge required to carry out problem-solving and to understand at what stage in a problem-solving process such knowledge is needed and their influence.

Thus, a learner can conduct appropriate decision-making by acquiring detailed knowledge based on this modular design method.

A series of cognitive activities are typical metacognition activities in PSOL. They include: a learner’s awareness of feasibility of a learning process as a start; monitoring one’s own knowledge state; comprehending its influence on a problem-solving plan; and building a learning plan for mastering knowledge required for problem solving. Reference to the appropriate information offered by a system to a learner encourages his or her appropriate metacognition activities, which help implement effective PSOL.

FUTURE TRENDS

Ontology-Aware System

The systems which support users to perform intelligent tasks based on the understanding of ontologies are called “ontology aware systems” (Hayashi, Tsumoto, Ikeda, & Mizoguchi, 2003). Systemizing ontologies contributes to providing theories and models, which are human-orientated to enhance systems’ abilities of explanation and reasoning. Furthermore, from the viewpoint of system development, building systems with explicit ontologies would enhance their maintainability and extendability. Therefore, future work in this field should continue developing systems that integrate ontology and HCI more effectively.

CONCLUSION

This article introduced a task ontology based human computer interaction framework and discussed various related issues. However, it is still difficult and time consuming to build high quality sharable ontologies that are based on the analysis of users’ task activities. Thus, it is important to continue building new methodologies for analyzing users’ tasks. This issue should be carefully addressed in the future, and we hope more progress can be achieved through collaboration between researchers in the fields of ontology engineering and human computer interaction.

REFERENCES


**KEY TERMS**

Attentional Capacity: Cognitive capacity divided and allocated to perform cognitive task.
**Metacognition**: Cognition about cognition. It includes monitoring the progress of learning, checking the status of self-knowledge, correcting self-errors, analyzing the effectiveness of the learning strategies, controlling and changing self-learning strategies, and so on.

**Ontology**: A specification of a conceptualization (Gruber, 1993).

**Problem-Solving Oriented Learning (PSOL)**: Learning not only to build up sufficient understanding for planning and performing problem-solving processes but also to gain the capacity of making efficient problem-solving processes according to a sophisticated strategy.

**Rasmussen’s Ladder Model**: A cognitive model that models human’s decision-making processes. This model is often used for human error analysis.

**Task Ontology**: A system of vocabulary/concepts used as building blocks for knowledge-based systems.

This work was previously published in the Encyclopedia of Human Computer Interaction, edited by C. Ghaoui, pp. 588-596, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 2.33
Social Networking Theories and Tools to Support Connectivist Learning Activities

M. C. Pettenati
University of Florence (IT), Italy

M. E. Cigognini
University of Florence (IT), Italy

ABSTRACT
This article considers the affordances of social networking theories and tools to build new and effective e-learning practices. We argue that “connectivism” (social networking applied to learning and knowledge contexts) can lead to a reconceptualization of learning in which formal, nonformal, and informal learning can be integrated as to build potentially lifelong learning activities to be experienced in personal learning environments.” In order to provide a guide in the design, development, and improvement both of personal learning environments and in the related learning activities, we provide a knowledge flow model highlighting the stages of learning and the related enabling conditions. The derived model is applied in a possible scenario of formal learning in order to show how the learning process can be designed according to the presented theory.

TOWARDS AN E-LIFELONG LEARNING EXPERIENCE
Formal, nonformal, and informal learning have become subjects of study and experimentation as for their potentialities to be carried on through the network. The pervasiveness of telematic technologies in current learning and knowledge processes justifies the hopes of success and emerging approaches become always more open, destructured, and nonformalised. According to this vision, formal, informal, and nonformal learning can be seen, such as integration of actions and situations, that can be developed both in the network and in
physical contexts. New reflections can therefore be made on the practice known as e-learning, starting from a revision of these dimensions. **Formal learning** has been defined as a type of learning that occurs within an organized and structured context (formal education, in-company training) and is intentional from the learner’s perspective. Normally it leads to a formal recognition (diploma, certificate) (Cedefop, 2000; European Commission, 2000). As regards adults e-learning, formal education in the last decade has encountered and experimented a sort of paradox that often witnessed low returns in terms of knowledge acquisition, compared to cost investment, which is often significantly high. **Nonformal learning** has been defined as learning embedded in planned activities that are not explicitly designated as learning, but that contain an important learning element. Nonformal learning is intentional from the learner’s point of view (Cedefop, 2000). **Informal learning** is learning resulting from daily life activities related to work, family, or leisure. It is often referred to as experiential learning, and can, to a certain degree, be understood as “accidental” learning. It is not structured in terms of learning objectives, learning time, and/or learning support. Typically, it does not lead to certification. Informal learning may be intentional but in most cases, it is nonintentional (or incidental/random) (Cedefop, 2000). Informal learning is an adaptive process determined by the exploration need, which is realised in specific experiential contexts (Calvani, 2005). People acquire their competence in everyday life, talking, observing others, trying and making mistakes, working together with colleagues more or less expert. Informal learning can therefore be intended as the natural corollary of daily life (Bonaiuti, 2006). **Intentionality** of learning is a discriminating factor shifting “nonformal” learning in “nonintentional” or “incidental” learning; contrary to what happens in formal learning, informal learning is not necessarily intentional and can be nonrecognized sometimes from the subject himself/herself as knowledge and competence acquisition (Cross, 2006). According to this perspective, aimed at retrieving and valuing the potentialities embedded in spontaneous contexts, in this case the network, the emerging domain of study of **informal e-learning** is receiving greater attention because of the widespread of social networking practices and technologies. The online transposition of the social network is nowadays referred to as “social networking” phenomena, and it is related to a set of available technologies and services allowing individuals to take part in network-based virtual communities. Social networking is emerging as a highly natural practice because it is deeply rooted in our daily practice; spontaneous relations, interactions, and conversations support informal learning practices, contributing to the creation and transmission of knowledge. In informal learning practices, the social behaviour and the support of technologies converge toward the “network”; a network made by people and resources, a social network, unified by personal needs or common goals, interaction policies, protocol and rules, and telematic systems all together favouring the growth of a sense of belonging to the “net” community.

At the same time, the culture of **lifelong learning** is gaining importance as one of the most effective answers to face the challenges brought by the information and knowledge society (Siemens, 2006): the rapid obsolescence of professional knowledge and skills requires updating and continuous training as well as recurring and personalised learning. Under these premises, the domain of e-lifelong learning is being configured as a sociotechnical system in which knowledge and learning are both the form and the content as for their social and relational meaning. The subject undergoing an e-lifelong-learning experience crosses this territory doing practices and strategies of continuous interconnection and com-
In this work the informal e-learning area constitutes a modularization of the educational sociotechnical system towards the full enactment of a lifelong-learning experience that could value the subject as for his personal and professional characteristics.

In 2004 the paper “Connectivism: A learning theory for the digital age” (Siemens, 2004) launched the theory of connectivism based on a critique of previous main learning theories synthetically labelled as behaviourism, cognitivism, and constructivism. According to Siemens, even the latter theory, which appeared to be the possible theoretical framework for e-learning practices (more specifically in its variant named “social constructivism”) could not provide an adequate theoretical support to the instances brought by the new learning approaches. To fill this gap, Siemens introduced the connectivism theory. According to this author, “The pipe is more important than the content within the pipe,” meaning that it is the network itself that is the basis of the learning processes. If the knowledge society requires the individual to continuously update his/her knowledge, this cannot happen as a process of progressive “knowledge accumulation”; instead, this can occur through the preservation of our connections. Learning, as it is also intended in Wenger’s community of practice vision (Wenger, McDermott, & Snyder, 2002) and in Lévy’s collective intelligence work (Lévy, 1996) is seen as a mainly social activity. In this perspective, the true competence for a lifelong learner of the knowledge society would be the capability to “stay connected” and “belong” to digital communities with which interests are and can be continuously shared. Owning a given piece of information is less important than knowing where and how to retrieve it.

Hereafter we will refer to e-lifelong learning as a possibility that each knowledge society citizen has to build: an individual and personalised lifelong-learning experience that will come across formal, nonformal, and informal learning stages with various degrees of uses of technology. In this domain, theoretical reflection and applied research is still at the beginning.

In this paper we try to provide a reference knowledge flow model to support the design of learning experiences in a networked environment accounting for new practices and technologies of social networking currently widespread in the Internet. The model aims also at giving suggestions to designers of personal learning and personal knowledge management environments in order to maximize the advantages deriving from the effective networking to enhance and improve learning and knowledge management functions.

To this extent, in paragraph 2, we analyse the affordances brought by the connectivist learning environment. In paragraph 3 we discuss the new buzz word of the domain e-learning 2.0 as a pretext to highlight methodology’s and technology’s role in learning. In paragraph 4 we introduce the conceptual view of a personal learning environment as a place where lifelong learning experience can occur. Then, in paragraph 5 we revise the social network technological applications that can serve the scope of the personal learning environment features. In paragraph 6 we discuss how learning can take place in such a context, introducing a knowledge flow model highlighting the learning stages and related enabling conditions in a connectivist environment. Eventually, in paragraph 7, we illustrate how the model of paragraph 6 can be applied in a formal learning situation of a possible higher education scenario.

AFFORDANCES OF LEARNING IN A CONNECTIVIST ENVIRONMENT

Scenarios, which become always more common, highlight that through informal channels, new learning and knowledge management spaces are more easily enabled, thanks to people and their
ability to “network” and reciprocally learn in a natural and spontaneous way. To synthetically analyse the strengths emerging in the context of learning in social network, we refer to Table 1. From this perspective, the main characteristics of social networks are schematized as for their sharing/cooperation/collaboration features.

The concept of social networks and their characteristics have been studied by Siemens (2006) as for their potential in knowledge management. This author first introduced the term of connectivism in order to make up for the lack of theoretical framework to learning in the digital age: “Research in traditional learning theories comes from an era when networking technologies were not yet prominent. How does learning change when knowledge growth is overwhelming and technology replaces many basic tasks we have previously performed?” (Siemens, 2006). “Knowing and learning are today defined by connections... Connectivism is a theory describing how learning happens in a digital age. Connectivism is the assertion that learning is primarily a network forming process” (2006). Siemens’ statement adds social characteristics to the territory in which learning (lifelong learning) can occur.

Connectivism is characterised by nine principles (2006):

1. Learning and knowledge require diversity of opinions to present the whole… and to permit selection of best approach.
2. Learning is a network formation process of connecting specialized nodes or information sources.
4. Knowledge may reside in nonhuman appliances, and learning is enabled/facilitated by technology.
5. Capacity to know more is more critical than what is currently known.
6. Learning and knowing are constant, on-going processes (not end states or products).
7. Ability to see connections and recognize patterns and make sense between fields, ideas, and concepts is the core skill for individuals today.

<table>
<thead>
<tr>
<th>Social Network Characteristics</th>
<th>Goal</th>
<th>Relation based on individual interests, debate, confront on specific topics; multiplicity and heterogeneity of joining interests and motivations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belonging</td>
<td></td>
<td>Spontaneous and autonomous motivation</td>
</tr>
<tr>
<td>Duration</td>
<td></td>
<td>Nondefined</td>
</tr>
<tr>
<td>Cohesion and enabling factors</td>
<td></td>
<td>High level of trust (relevance of reputation), sense of responsibility, high technological skills, distributed reflexivity and evaluation (nonautonomous, nor heteronymous but socially spread) Type of relation: share/evaluate</td>
</tr>
</tbody>
</table>

Table 1. Social network characteristics (Adapted from Pettenati & Ranieri, 2006b)
8. Currency (accurate, up-to-date knowledge) is the intent of all connectivist learning activities.
9. Decision making is learning. Choosing what to learn and the meaning of incoming information is seen through the lens of a shifting reality. While there is a right answer now, it may be wrong tomorrow due to alterations in the information climate affecting the decision.

According to these principles, learning, as it was already conceptualised in Wenger et al.’s (2002) community of practice vision and in Lévy’s (1996) collective intelligence theory, is seen as a prevalently social activity. “Connectivism is the integration of principles explored by chaos, network, complexity, and self-organization theories. Knowledge and learning are processes that occur within nebulous environments of shifting core elements—not entirely under the control of the individual. Learning, defined as knowledge patterns on which we can act, can reside outside of ourselves (within an organization or a database), is focused on connecting specialized information sets” (Siemens, 2006). Connectivism is therefore social networking applied to learning and knowledge.

A New Role for the Learner: Continuously Change and Be Creative

If from one side we value connectivism as a context in which learning can more favourably occur, thanks to available technological solutions (Fallows, 2006), on the other side we acknowledge that connectivism is also enabled and allowed by an always stronger user participation to the creation, sharing, use, and management of resources (contents, relations, applications, etc.) through social software. Users have certainly become more aware consumers and producers of resources; they have become prosumers.

This radical change of the users’ role is under our eyes; between 2000 and 2004, relevant literature reported about research investigating the possible use of blog tools in education without being capable of bringing credible answers. The use of the tool at that time had not yet entered in the praxis and communicative modality of the people and the potentialities related to the aspects of metareflection and self-evaluation of the blog had not yet clearly emerged (Barrett, 2004; Barret 2006; Banzato, 2006). More recently, the exponential growth of blog use, coupled with the syndication technologies, which added the “relational” dimension to these tools, made the blog one of the most important social networking assets used at the moment. This is just an example that serves to highlight that the determinant variable of the shift in the use of the tool was not due entirely to technology, but also to the spontaneous change in the practice of use together with the overcoming of the diffusion critical threshold that could make the tool an important instrument in learning (formal, informal, nonformal) processes. Another relevant mark of this change in the users’ role is related to recent news; in December 2006, all the blogosphere rumoured about TIME magazine’s cover title: “TIME’s Person of the Year for 2006 is you.” According to Lev Grossman (2006), author of this news, the explosive growth and the enormous influence of the user-generated content (such as blogs, video sharing sites, etc.) should be read in a new way: “But look at 2006 through a different lens and you’ll see another story, one that isn’t about conflict or great men. It’s a story about community and collaboration on a scale never seen before. It’s about the cosmic compendium of knowledge Wikipedia and the million-channel people’s network YouTube and the online metropolis MySpace. It’s about the many wrestling power from the few and helping one another for nothing and how that will not only change the world, but also change the way the world changes.”
“E-LEARNING 2.0” OR JUST “LEARNING”?

Innovation in technology has brought new (or revisited) processes and practices, expressed through new (or revisited) criteria and terms: the possible educational universe of e-learning, formal, informal, nonformal, lifelong coupled with connectivism features has been named (or tagged) e-learning 2.0 or learning 2.0, analogous to what happened for the Web 2.0 phenomena (O’Reilly, 2005). E-learning 2.0 is therefore a tag identifying an open, destructured, immersive, and relational learning process amplifying the learning curve towards the social, relational side of knowledge coconstruction over the net. Stephen Downes (2005) illustrated the main concepts of e-learning 2.0:

1. it is a type of learning that is tied to the interests of the learners,
2. it is an immersive learning that is acquired by doing, realised not only in the class, but also in appropriate environments (for instance museums...),
3. it is a type of learning in which Web 2.0 technologies connect the learners to the rest of the world.

However stated, the focus stays on the attention risen by the affordances that the social-technical connectivist learning environment can engender, in light of a background that remains primarily and unavoidably methodological: the design and development of educational environment is, beyond any revisiting, a learner/learning-centred process.

When knowledge technologies change so radically, they change not only “what” we know, but also “how” we come to know (Laurillard, 2003). Knowledge is the result of a fluid combination of experience values, contextual information, and specialist competences, all together providing a reference framework for the evaluation and assimilation of new experiences and knowledge (Pettenati & Ranieri, 2006b). Norris, Mason, and Lefrere (2003) highlights the recursive, dynamic, and networked character of learning in digital contexts: “Knowledge can be understood as interpreted content, available to a member of a community and always shaped by a particular context. Digital representations of content and context become e-knowledge through the dynamics of human engagement with them. The digital elements of e-knowledge can be codified, combined, repurposed, and exchanged.” The relational aspect of learning and knowledge processes that is proposed in this work is supported by Wilson’s (2002) essay “The nonsense of knowledge management,” in which the author states that: “‘Knowledge’ is defined as what we know: knowledge involves the mental processes of comprehension, understanding and learning that go on in the mind and only in the mind, however much they involve interaction with the world outside the mind, and interaction with others. Whenever we wish to express what we know, we can only do so by uttering messages of one kind or another oral, written, graphic, gestural or even through ‘body language.’ Such messages do not carry ‘knowledge,’ they constitute ‘information,’ which a knowing mind may assimilate, understand, comprehend and incorporate into its own knowledge.” Analogously, we attribute to e-learning a social connotation, for us, e-learning is a type of learning that is somewhat supported by technologies, but it is not necessarily conducted at a distance; it allows interaction between people and contents, and among people but (most important) is a type of learning valuing the social dimension of the underlying knowledge processes (based on Calvani, 2005) definition, freely translated and adapted by the authors). The presence of the “e-” before all learning and knowledge concepts is therefore now being abandoned by many researches to underline that pervasiveness of technologies do not differentiate any more “traditional” contexts from “virtual”
Social Networking Theories and Tools to Support Connectivist Learning Activities

ones. The “traditional” concepts of learning and knowledge, newly restated and deprived by the Internet-ic “e-,” globally revolutionize the education and knowledge systems, both in the practices and in the social processes they imply. As the learner in an early stage of the knowledge society had to change and shift identities and practice to access the same processes over the network (Liguorio & Hermas, 2005), so the learning and knowledge practices regain possession of their integrity to come back to their original complexity, now completed and evolved by technologies but not distorted by the same technologies.

**Personal Learning Environments as New Learning Landscapes**

In this scenario, many of the innovative technolo-
gies of the so-called “social software” (see next paragraph) are now playing a crucial role as a support to learning and knowledge processes. This does not mean that the scaffold provided up to now by a formalised online educational envi-
ronment is to be entirely rebuilt, but the current “extended cognitive context” (environmental, situational, cultural characteristics) in which the learning dynamics occur must reshape the learning environment itself (Bonaiunti, 2006). Metaphorically getting out from the often narrow space of the learning/course/content management systems is equivalent to the reasoning conducted up to now: learning does not take place only in classes, for a limited period of time, and under specific certification conditions but (in light of the sociotechnical context now available) can easily take place in nonformal, informal, lifelong settings, prospecting new, challenging and interesting development perspectives. Most evidently, according to this perspective, new coordination and integration needs emerge, as for the experiences conducted in heterogeneous systems, from the functional and from the pedagogical and methodological viewpoints. A new technological solution is never a didactic solution, Quite the opposite, it urges to be considered after the methodological design has been carried out, in order to properly balance the use of the tools with respect to learning objectives, learners, available resources, and infrastructures (Ranieri, 2005). Stating that the “social software” (or Web 2.0) technological apparatus has great educational potentials would be reductive and misleading; which type of learning can occur and under which conditions and ways this learning can oc-
cur should always be methodological priorities to be accounted for during instructional design. The shift required at this stage of the reasoning can be conducted with the avail of a specific environment, personalised on the needs and the competences of the learner, which could always be a “door” to enter learning experiences and a “glue” to support the learner in mastering the heterogeneity of his learning experiences, easily jumping across formal, informal, unexpected, and intentional learning activities (Conner, 2004). This “door-glue-environment” is represented in the hopes of the practitioners of this domain, and in concrete available solutions, by personal learning environments (PLE) that are a “concretization of operational spaces in the network where the subject is at the centre of his network of learning resources” (Bonaiuti, 2006).

PLEs represent the temporal horizon through which we look at the potential lifelong learning of a subject (Tosh, 2005; Tosh & Werdmuller, 2004); they are an open, transversal, and personal environment intercepting and capitalizing learners competences and skills by aggregating, publishing, combining, and integrating diverse and distributed resources and contents. E-portfolios are another way to name PLEs (e-Portfolio, 2006); “an e-portfolio can be seen as a type of learning record that provides actual evidence of achievement. Learning records are closely related to the Learning Plan, an emerging tool that is being used to manage learning by individuals, teams, communities of interest and organizations.” The attention given recently
to this topic at the eStrategy conference on e-Portfolio (Baker, 2006) is but another evidence of the rising importance of studying this field. “In the context of a knowledge society, where being information literate is critical, the portfolio can provide an opportunity to demonstrate one’s ability to collect, organise, interpret and reflect on documents and sources of information. It is also a tool for continuing professional development, encouraging individuals to take responsibility for and demonstrate the results of their own learning. Furthermore, a portfolio can serve as a tool for knowledge management, and is used as such by some institutions.”

Technology to Support the Learning Experience

The further step in the analysis leads us to the problem of evaluating and devising which tools and technologies exist or can be developed in order to match the requirements and purposes of the personal learning environment. Technologies and tools now referred to as Web 2.0 software (Fallows, 2006; Hinchcliffe, 2006; O’Reilly, 2005) certainly provide both the origin of this reasoning as well as the goal at which to aim. Sharing Paul McFedries (McFedries, 2006) tentative definition according to which Web 2.0 is “a second phase of the evolution of the World Wide Web in which developers create Web sites that act like desktop programs and encourage collaboration and communication between users,” we focus on the characteristics of the Web 2.0 applications, highlighting the social perspective of relation, collaboration, and user-participated architecture (McFedries, 2006):

1. content is user-created and maintained (peer production, user-content ecosystem),
2. user-created and maintained content require radical trust,
3. application usability allows rich user experience,
4. combining data from different sources leads to creation of new services,
5. services get better as the number of users increase in an architecture of participation.

WEB 2.0 TOOLS

Folksonomies, cobrowsing, tagging, and social networking are “2.0” practices. The fil rouge they share is that they all are expressions of a shared, diffused cognitive strategy for information retrieval in a spontaneous way, as support to social sharing tools (such as social bookmarking tools, image sharing tools, blog search engines, etc.). Through social tagging, the member of the community defines a link among resources (sites, images, videos, audios, etc.) and the terms used to describe them (Bonaiuti, 2006). This is a bottom-up process, starting from a single user adding a link to a Web site and tagging it at his complete discretion, using keywords that are meaningful to himself/herself. Social sharing tools can display these tags through using a visual approach (which increases the font size of most popular tags), thus realising tag clouds that immediately provide users with a perception of the popularity of the tags. This “folksonomic” classification method, which relies on the spontaneous users’ contributions (be he/she an author or a sporadic resource user), leads to results capable of reflecting the information according to the conceptual model of the population that creates it (Bonaiuti, 2006). Representing information in classificatory structure is information in itself: classification incorporates information and provides the interpretation context that (in its globality) appears transparent, objective, and neutral (Surowiecki, 2004).

Practices and applications described as an example implement ideas derived from “social network theory,” which sees social relations expressed through nodes and links in which nodes represent the subjects in the network and
the links represent relations between subjects. Under this theoretical basis, these applications explicitly represent relations favouring the interconnection among subjects based on interest, competence, hobby, research goals, study, work, and so forth (Fini, 2006). Social networking is

**Table 2. Sample Web 2.0 applications, a synthetic view**

<table>
<thead>
<tr>
<th>Web application</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social networking, online social networks</td>
<td>Category of Internet applications to help connect friends, business partners, or other individuals together using a variety of tools.</td>
</tr>
<tr>
<td>Social network search engines</td>
<td>Social network search engines are a class of search engines that use social networks to organize, prioritize, or filter search results.</td>
</tr>
<tr>
<td>Blogs</td>
<td>A Web log is a Web site where entries are displayed in chronological order. They often provide commentary or news on a particular subject, typically combining text, images, and links to other blogs, Web pages, and other media related to the specific topic.</td>
</tr>
<tr>
<td>Blog guides</td>
<td>Specialized search engines for searching blog and news contents.</td>
</tr>
<tr>
<td>Social tagging (folksonomy)</td>
<td>Ad hoc classification scheme (tags) that Web users invent as they surf to categorize the data they find online.</td>
</tr>
<tr>
<td>Social bookmarking</td>
<td>Saving and applying keywords to one’s personal collection of Web site bookmarks on a site that enables other people to share those bookmarks.</td>
</tr>
<tr>
<td>Web syndication, Web feed management</td>
<td>Web syndication is a form of syndication in which a section of a Web site is made available for other sites to use through to making Web feeds available from a site in order to provide other people an updated list of content from it (for example one’s latest forum postings, etc.).</td>
</tr>
<tr>
<td>Tag clouds</td>
<td>A list of tags used in the site with some kind of visual indication of each tag’s relative popularity (ex. large font). Web sites that implement tag clouds functions allow both finding a tag by alphabet and by popularity. Selecting a single tag within a tag cloud will generally lead to a collection of items that are associated with that tag.</td>
</tr>
<tr>
<td>Peer production news</td>
<td>Web sites combining social bookmarking, blogging, and syndication with a form of nonhierarchical, democratic editorial control. News stories and Web sites are submitted by users, and then promoted to the front page through a user-based ranking system.</td>
</tr>
<tr>
<td>Wikis</td>
<td>Collaborative Web sites that allow users to add, edit, and delete content.</td>
</tr>
<tr>
<td>Collaborative real-time editing</td>
<td>Simultaneous editing of a text or media file by different participants on a network.</td>
</tr>
</tbody>
</table>

*continued on following page*
the connective tissue in which new applications emerge daily, allowing us to creatively use them for learning purposes. The following table reports a synthetic view of some of the currently available applications matching this philosophy; however, the list is neither complete nor exhaustive, and is intended just to be a memorandum for personal learning environment designers.

**A MODEL FOR EFFECTIVE LEARNING EXPERIENCE**

To comply with the objectives detailed in the previous paragraphs, we hereafter propose a schema to illustrate a (simplified) knowledge flow in a connectivist environment. This model is the result of a re-elaboration of four models:

- the online collaboration model presented in Calvani (2005), which was the starting point of our analysis; this model accounts for effectiveness conditions and principles that are considered to be fundamental for collaboration, as highlighted in reference literature (Dillenbourg, Baker, Blaye, & O’Malley, 1996). However, while in Calvani (2005) the model is conceived to provide useful steps to support an online collaborative group, the model is rooted in a formal educational context;
- the second model studied is the one presented in Ranieri (2006); this model accounts for individual and collaborative knowledge construction processes, but it is not specifically conceived focused on telematic environments, and does not account at all for collaboration enabling conditions;
- the third and fourth models are those presented in Pettenati and Ranieri (2006a, 2006b) in which the authors tried to focus on framing the reasoning in a social networking context in order to account for the benefits of informal learning and online collaboration, as described in the previous paragraphs of this paper. Nonetheless, that model did not account for the knowledge flow (or learning stages) occurring in an online learning experience, thus being only partially helpful in the design of a learning experience.
After having confronted the previous three models with the knowledge flow presented in Siemens (2006), the idea of the model was repurposed in light of the “connectivist” idea of the author’s work, thus leading to a new design of the model now focused on highlighting the possible stages on an effective connectivist learning experience. In this new version, the schema can provide more concrete help in the design of online learning activities, as will be detailed in section 7, using an example. In Figure 1, the transformation processes that involve personal comprehension and knowledge as well as social knowledge creation are illustrated in order to highlight some important processes that should occur during an effective learning experience. Rectangles in Fig. 1 represent the knowledge processes whose results can be concretized in various forms of knowledge (not illustrated in Figure 1) such as enunciations, argumentations, justification, shared meanings, collaborative knowledge, cultural artefacts, and so forth (Ranieri, 2006).

The schema in Figure 1 has no claims to be exhaustive neither perfectly accurate, as knowledge processes are complex, fluid, iterative, and certainly not univocally directed, thus leading to various relations among the composing elements (represented by rectangles in Figure 1). However this schema will be used in section 7 to drive the choice of methodologies and available technologies and their combination in order to serve the educational needs in a possible higher education scenario.

**Enabling Conditions**

The model in Fig. 1 envisages five subsequent stages (or knowledge processes) that are at the heart of the schema. The processes are framed by an external layer, where the enabling conditions that are relevant for the knowledge processes development are highlighted:

*Figure 1. Knowledge process in a connectivist environment: stages of the learning experience and enabling conditions*
1. **Basic skills:** The acquisition of basic technological as well as online-comunicational skills is fundamental (Ranieri, 2006) to allow the subject to positively enter and live the informal and potentially lifelong educational experience; the lack of these conditions can compromise the success of the entire learning initiative;

2. **Generation and support to motivation:** In informal e-learning contexts the motivation is spontaneous; it is often induced by fun and pleasure that individuals have in their network activity; it is also rooted in the positive interaction among people (a subject can more effectively and efficiently pursue his objective if the other subjects pursue theirs);

3. **Meaning perception:** The subject must perceive as really meaningful (useful to himself/herself) the objectives attainable in the learning activities and acknowledge that collaboration can derive real advantage; during the collaboration phases, supporting the individual self-perception of usefulness allows the subject to perceive the significance of its contribution to group activities in order to consider himself a useful contributor to other’s goals;

4. **Group culture:** The awareness of being useful to other community members increases the self-esteem and self-perception of usefulness of the individual and fosters the motivation for a wider visibility (for instance being linked, have positive reputation, produce and/or propose new contents). These factors increase the sense of positive group membership;

5. **Social climate:** In an informal environment, the sense of belonging (membership) to a group is spontaneously supported by the intensity of sharing interests on a topic; regardless of the expertise (which can be widely nonhomogeneous among members) it is still the awareness of the positive interaction with others that sustains mutual understanding and social grounding; in this context the (often tacit) agreement of respect, use of reputation feedback, and respect of a common socioquette, contribute to build a positive social climate, making the online relational environment a “trusted” environment.

### Stages of the Learning Experience

In the centre of Figure 1, a simplified schema of the stages of a connectivist learning experience are schematized:

1. **Awareness and receptivity:** Siemens (2006) points out, in this stage, individuals get used to “handling knowledge abundance,” and are first confronted with resources and tools of the new learning habitat. If this stage is not sustained by a proper motivational context as well as the acquisition of basic skills, as detailed in paragraph 5.1, this phase can become a frustrating experience and often cause the learner to drop out from the learning environment.

2. **Connection forming and selection filtering:** In this stage, individuals begin to use tools and understanding acquired during the previous stage to create and form a “personal network” of resources (people and contents). At this point the learners start to be active in the learning space in terms of “consuming or acquiring new resources and tools” (Siemens, 2006). Affective and emotive factors such as fun, pleasure, positive interaction, and increasing sense of meaning are crucial to building the roots of an effective “personal learning network” (ibidem).

3. **Contribution and involvement:** “In this stage the learner begins to actively contribute to the learning network, essentially,
Social Networking Theories and Tools to Support Connectivist Learning Activities

becoming a ‘visible node.’ The learner’s active contribution and involvement allows other nodes on the network to acknowledge his/her resources, contributions, and ideas, creating reciprocal relationships and shared understandings (or, if social technology is used, collaboratively-created understandings)” (ibidem). In this stage group culture and social climate are key factors enabling the individual participation and involvement.

4. **Reflection and Metacognition:** Reflection on the knowledge processes and products, self-reflexivity, and self-evaluation as well as metacognition (thinking about thinking) play a prominent role in this stage. Individuals are actively involved in modifying and rebuilding their own learning network, acting as “network aware and competent” subjects (ibidem). The experience acquired at this stage within the learning network has resulted in an understanding of the nuances of the space and the knowledge inputs, allowing the subject to act both as a provider of valuable support and to help other networked learners as well as being capable of accessing just in time and personalized knowledge to himself/herself.

The learning process depicted can be used as a model to support design of learning activities and environments in different educational sets. An example of this approach will be shown in paragraph 7 to build a “connectivist learning experience” in a possible higher education scenario.

**Building a Connectivist Learning Experience: A Possible Scenario**

This scenario subject is taken by Bonaiuti (2006), but is significantly detailed, revised, and newly implemented according to what is presented in this paper. More specifically, the focus is on building and exploiting a connectivist online environment according to the model in Figure 1, to match the educational purposes.

**Scenario:** Higher education, course in “Urban Sociology,” class topic: “Juvenile behaviour in urban suburbs” (Bonaiuti, 2006). Scenario details: this course envisages 2 hours face-to-face class time and a total of 10 hours to be devoted to autonomous and group work. The teacher wants to structure the face-to-face meeting as a “role playing” activity. The scenario for the role play is the following: participants: a math teacher, a chancellor, two parents, a young boy. The young boy has been caught by the math teacher outside of the school while cutting the teacher’s car wheels. The teacher has reported the happening to the chancellor and the student is now risking expulsion from the school. Debate with arguments in favour of expulsion, neutral, in defence of the student, and reflective, will be played in 1 hour during the face-to-face class meeting.

The model of Figure 1 can be used to help the teacher settle his/her teaching activities. To this extent, the model is exploited during the three phases of class activities: preclass activities, in class activities, postclass activities. Instrumental background: the university in which this scenario is set has installed a learning environment at the disposal of teachers and students who want to avail of it. This environment is a combination of a “traditional” e-learning environment and a “personal learning environment” conceived according to the e-learning 2.0 philosophy, as detailed in the previous part of this work, and opened to students and alumni who want to keep alive their online lifelong learning landscape. Such a combined environment offers both the course and content management system offered by traditional e-learning systems coupled with social-networking creation, sharing, and management tools.
1. Preclass Activities: Focus on Content Management and Sharing

a. **Awareness and receptivity**: teacher gives starting tasks to the student using the blog of the learning environment in the area reserved for the Urban Sociology course. In this post, he/she proposes a selected reference reading (full text of the reading could be stored in the Course Material area) and suggests a vision of three short videos distributed through a video-sharing social network. The assignment for this phase is to prepare the debate that will take place during the face-to-face class: students are divided in five groups; each group is charged to represent one role in the scenario: math teacher, chancellor, two parents, young boy. Each group, in order to get prepared to sustain the arguments and debate, can find external documentation to get closer to the discussed reality of juvenile behaviour in urban suburbs (multimedia resources are allowed, such as podcast, video, text, etc.).

b. **Connection forming, selection, filtering**: in this phase students prepare themselves to work in a group; each group decides how to carry on its task (for instance attribute roles to the group members, share tasks, provide complementary resources, etc.). This work is done in spontaneous modes for each group: some can use the course forum; teacher can open related threads, or can work in synchronous mode using voiceIP tool and sharing ideas and outline, or to dos using the class wiki or a conceptual map tool; teacher can provide a tool area for each group of the course.

c. **Contribution and involvement**: in this phase students actively work using tools and combining resources to come to a shared strategy of discussion for the class meeting. Students post the outline of their argumentations in the course blog or forum (and can tag it with the group name). Students can use the social bookmarking feature of the course environment to organize and share their resources.

2. In-Class Activity: Focus on Content Production

a. **Contribution and involvement**: during the first hour of the class the role-playing game is set. The teacher listens to the debate and uses a mental map tool to synthesize the emerging discourse pattern.

b. **Reflection and metacognition**: after the end of the play, the teacher restates what happened and comments on group contributions. He/she provides ideas and questions for further insights and refrares all the play in the context of the topic of the class, with reference to the selected reading assigned before the class. At the end of the class, both the conceptual maps drawn and the audiorecord of the play are shared on the course environment.

Students are then given the evaluation assignment to be carried in a post-class phase within 15 days. The assignment is a journalist-style individual writing of a possible local newspaper reporting the case of the role play in light of the analysis of the juvenile behaviour in the urban suburbs. Students are asked to post the assignment in their personal blog, tagging it with the course name and class topic.

3. Post-Class Activities: Focus on Knowledge Acquisition

a. **Contribution and involvement**: students will work individually to study and analyze both the material produced by the whole class and the material provided by the teacher. Students can work at the writing using either an online (to have easy backup and document
portability) or desktop word processor, or a wiki (to keep trace of the evolving versions). Eventually they will post the assignment in their personal blog, tagging it as required. The teacher will easily follow the assignment submissions using a syndication system.

b. Reflection and metacognition: to support the production phase the teacher (or a tutor acting on his/her behalf), in the meanwhile, will use the course blog to more deeply comment on the topic or provide further resources, as well as posting comments on students’ posts (both on the blog, or forum or wiki). After the evaluation phase (marks have been given to student’s assignment in a privacy-respectful way), the teacher could invite students to “digg” their preferred writing to highlight them in the school journal.

CONCLUSION

In this article we tried to provide our interpretation of the current sociotechnical educational system shaped by technologies and practices of the “knowledge society” to locate the role of learning and learners in a lifelong perspective. We believe that both users’ attitudes and available technologies are mature enough to let us envisage that each network user could easily engage in a lifelong learning personal experience if properly lead by appropriate methodologies, and sustained by accordingly designed and developed personal learning environments.

To this extent we provided a model to schematize the knowledge flow occurring during an effective learning experience in a connectivist environment. The purpose of this model is two-fold: from one side it can be used by personal learning-environment designers as a guideline for checking if all phases and enabling conditions are supported by the integrated tools; on the other side it can be used by instructors or designers to set up learning activities. The provided model has some common points with other models at the state of the art; for instance in Salmon (2002), the author provided a staged model for online collaboration activities oriented to knowledge construction; however, that model, and the related implementations and example presented in the essay, was developed in the “1.0” age, where neither technology nor people practice was mature enough in the sense of social networking perspective. We believe that a connectivist revision of this theory is now necessary, and this work intended to be a contribution in this sense. As a conclusion of the paper, we provided a sample application of the model using the scenario. The scenario setting chosen for this purpose is a formal learning higher education context; this choice should not be read as contrasting with the lifelong perspective of learning claimed throughout this work. Instead, it should be read as a preparatory activity that could enter in people’s practice diffused by a formal context but (once mastered as skill) potentially applied in all other nonformal and informal learning contexts.

Applying the knowledge processes flow model in other scenarios such as professional training in enterprises, as well as refining and detailing the model to these purposes, could be an object of future study. Many other issues are still open for further investigation and experimentation: how could a personal learning environment be used to develop pedagogical thinking? How can they be used in professional contexts such as workers for placements/reintegration?, How can the formal educational system be changed as to support self-directed learners?, What arguments can convince politicians (educational and labour ministries) to support individual institutions? These questions, as many others, were not addressed in this work, but can and will be object of future studies. Whether we call it e-learning, e-learning 2.0, lifelong learning, knowledge management or not, the interesting question is to see if and how these approaches can lead us to a true actualization
of a lifelong learning practice for all knowledge society members.

ACKNOWLEDGMENT

We want to thank Prof. Dino Giuli for giving us the possibility to carry on our research in this domain. We are also very grateful to Prof. Antonio Calvani for keeping on asking us stimulating scientific research questions, pushing us to work in order to find possible answers. Moreover, we want to thank the colleagues of the Educational Technology Laboratory and Telematics Technologies Laboratory for the fruitful discussions occurring both in formal and nonformal settings.

REFERENCES


Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. Little, Brown.


Chapter 2.34
User-Centered Design Principles for Online Learning Communities: A Sociotechnical Approach for the Design of a Distributed Community of Practice

Ben K. Daniel
University of Saskatchewan, Canada

David O’Brien
University of Saskatchewan, Canada

Asit Sarkar
University of Saskatchewan, Canada

ABSTRACT

This chapter examines current research on online learning communities (OLCs), with the aim of identifying user-centered design (UCD) principles critical to the emergence and sustainability of distributed communities of practice (DCoPs), a kind of OLC. This research synthesis is motivated by the authors’ involvement in constructing a DCoP dedicated to improving awareness, research, and sharing data and knowledge in the field of governance and international development. It argues that the sociotechnical research program offers useable insights on questions of constructability. Its attention in particular to participatory design and human-computer interaction are germane to designing user-centered online learning communities. Aside from these insights, research has yet to probe in any systematic fashion the factors affecting the performance and sustainability of DCoP. The chapter concludes with a discussion of UCD principles for online learning community to support the construction and deployment of online learning communities.
INTRODUCTION

Increasingly, distributed communities of practice (DCoPs) are attracting attention for their potential to enhance learning, to facilitate information exchange, and to stimulate knowledge creation across cultural, geographical, and organizational boundaries. Research shows the utility of DCoP on their members is positive (Daniel, Sarkar, & O’Brien, 2004a; Daniel, Poon, & Sarkar, 2005; Schwier & Daniel, Chapter II, this volume). Their allure aside, experience indicates that they may not emerge or flourish even in the presence of demand from users. In fact the process of constructing DCoP is not well understood, and factors influencing sustainability merit further research attention.

This chapter introduces the authors’ involvement in the development of a DCoP. The DCoP in question is the Governance Knowledge Network (GKN). This project began in 2001 with the aim of assessing the interest of academics and practitioners in Canada to develop an online learning community (OLC) for systematizing the exchange of information at the intersection of governance and international development (Daniel et al., 2004a). The surveys of key Canadian stakeholders in the project indicated considerable data existed, and recommended the proposed GKN to: actively engage in dissemination and archiving of data not widely accessible in the public sphere, profile community members, promote social network building and collaboration, and inform members of current events and opportunities.

Following the identification of the demand and interest, the second stage of our research involved the development of a GKN prototype. In this unchartered course, we were guided by enabling technology and other DCoP models (World Bank, UNDP). We also turned to research to inform our efforts on how to effectively sustain the project. Our synthesis of research in the area identified promising insights from studies we refer to as the sociotechnical approach. As applied to DCoP, the sociotechnical approach aims at understanding people’s interaction with technology and the ensuing communication, feedback, and control mechanisms necessary for people to take ownership of the design and implementation process.

This chapter focuses on this interaction, as it is germane to the development and sustainability of the GKN, in particular, and DCoP more generally. The chapter is divided into the following sections. The next section outlines relevant research on DCoPs and the sociotechnical approach. We next provide an overview of the GKN OLC project and present key results from the research that informed the design of the GKN. A discussion of various human and technology elements we consider critical to the initiation, development, growth, and sustainability of the GKN follows, and in the next section, we revisit the key human and technology design issues. Finally, we conclude the chapter and present UCD principles for OLCs drawn from the sociotechnical approach.

RELATED WORK

Daniel, Schwier, and McCalla (2003b) observe that online learning communities have attracted diverse disciplinary interest, but that it is possible to identify two dominant perspectives—technological determinism and social constructivism. The basic tenet of the technology determinism research is that technology shapes cultural values, social structure, and knowledge. Social constructivism theories have inspired research on knowledge construction within communities of practice. Lave and Wenger (1991) assert that a society’s practical knowledge is situated in
relations among practitioners, their practice, and the social organization and political economy of communities of practice. For this reason, learning should involve such knowledge and practice (Lave & Wenger, 1991). Between these heuristic poles there are cross-disciplinary perspectives, of which it is possible to further discern them into four subcategories:

1. **Applied Technology Perspective:** Much of the work on OLC by computer scientists and information systems researchers is driven by a desire to understand and improve computational approaches. Studies in computer science, information systems, and educational technologies are mainly aimed at understanding technology to develop tools and systems that support learning environments (Daniel, Zapata-Rivera, & McCalla, 2003a; Preece, 2002; Schwier, 2001). Findings have been utilized for building technologies that support OLC. For instance, a growing number of developers and researchers in industry and universities are investigating ways to create software packages that add new functionality to systems supporting interaction, collaboration, and leaning in online learning communities (Kim, 2000; McCalla, 2000; Preece, 2000; Resnick, 2002; Schraefel, Ho, Milton, & Chignell, 2000).

2. **Ethno-Narrative Perspective:** Ethno-narrative research is devoted to revealing personal experiences of being a member of an OLC. Most studies adopt a narrative approach, similar to participant observation inquiry used in anthropology. Researchers in this tradition have undertaken comparative analysis of both online learning and temporal communities (Schwier, 2001). Critics have disparaged ethno-narrative studies on the grounds that findings tend to be anecdotal and lack external validity; their conclusions are tentative and limited to the groups under study (cf. Downes, 2001; Rhiengold, 1993, 1999, 2002). Stolterman, Croon, and Argren (2000) argue that although the generalization and validity of such studies is limited, understanding personal perceptions of learning in OLC is essential. It is difficult to imagine how one can improve the learning environment of OLC without the subjective feedback of the learners.

3. **Cultural Studies Perspective:** Cultural studies have contributed enormously to understanding online learning communities. For instance, research by Brook and Boal (1995), Dery (1994), and Hershman and Leason (1996) investigate the relationship between the virtual and the physical, and they fall within the context of cultural interpretation research. Approaches employed in this category include experimental studies, with an emphasis on cultural events in online environments. The background disciplines of this group are diverse, including social psychology, philosophy, psychology, and fine arts.

4. **Sociotechnical Perspective:** The sociotechnical research tradition argues for a balanced approach to integrating cognitive and technical dimensions of OLC. This approach emerged from the extension of sociology, anthropology, and psychology to the study of HCI. Subsequently this research informed disciplines, including computer science and information systems (Heylighten, 1999). Research in sociotechnical areas addresses issues such as:
   - **User-Centered Design:** Moving the focus of interest to learners and away from technology in the design of online learning (Norman, 1996).
   - **Contextual Enquiry:** Understanding the user’s context and its potential influence on the use of technology (Preece, 2000).
   - **Sociability:** Appreciating the importance of community policies for
interactions, governance, and social protocols in OLC (Preece, 2000).

- **Participatory Design**: Involving user participation in the design of OLC and the effects on learning outcomes (Mumford, 1987; Nguyen-Ngoc, Rekik, & Gillet, Chapter XIII, this volume).

- **Direct-Manipulation**: Creating tools for users to create their online learning environment and exploring the effects of functional options such as menu-driven and graphical interfaces (Shneiderman, 1998).

Common to this growing body of research issues is the need for the interplay of human and technology factors to guide the design, development, deployment, and evaluation of online learning communities.

**Formal and Informal Online Learning Communities**

There are numerous computational tools that support social learning across time and place (Laghos & Zaphiris, Chapter XI, this volume). New tools and patterns of communication have enabled social engagement, information, and knowledge sharing within social systems now referred to as OLC. Unlike a temporal community that resides in a fixed locale and whose members often know each other well enough to carry effective interactions, OLCs exist in cyberspace and may or may not be aware of each other (Daniel, Schwier, & McCalla, 2003). The character of an OLC is influenced by structural features, which may include: community size, duration of interaction and anticipated lifespan, location or distribution of the community, the homogeneity/heterogeneity of members, and breadth or narrowness of subject area. Variation of these features gives rise to diverse OLCs.

In Table 1, we simplify this diversity by distinguishing between formal and informal online learning communities. Formal online learning communities have explicit learning goals and evaluation criteria. Examples would include courses/programs offered by education institutions or companies (McCalla, 2000; Schwier, 2001). In contrast, informal OLCs achieve learning outcomes through social learning. Examples would include distributed communities of practice (Daniel, O’Brien, & Sarkar 2004b). A unique feature of DCoPs is the absence of a teacher or instructor; rather, in a DCoP, the learners are also teachers, as members collectively determine the content and support each other throughout the learning process. Further differences are contrasted in Table 1.

A growing body of research identifies the contribution of DCoPs to facilitating information exchange and knowledge creation, thereby enriching the work of the collective (Brown & Duguid, 1991; Hildreth, Kimble, & Wright, 1998; Lesser & Prusak, 2000). These positive outcomes have caught the interest of scholars and knowledge managers. And yet, there is little comparative research on the correlates of DCoP performance or sustainability. We find this surprising, given the fact that OLCs emerged and proliferated with the advent of the Internet and then World Wide Web over a decade ago. The case-study foundations for comparative research are certainly present, however (Kalaitzakis, Dafoulas, & Macaulay, 2003; Hartnell-Young, McGuinness, & Cuttance, Chapter XII, this volume).

Germane to the topic of DCoP emergence and sustainability is the question of “constructability”. Can the DCoP features listed in Table 1 be built, or have DCoPs simply migrated from the temporal to the online world? If we return to the literature review briefly touched on earlier, perhaps not surprisingly we would find a different answer to this question depending on the literature consulted. For example, the sociology
and cultural studies literature tends to be skeptical of the view that DCoPs can be constructed (Kollock & Smith, 1996). By contrast, the computer science and information systems research, on the whole, seem more optimistic that robust DCoPs can be constructed (Preece, 2000; Daniel et al., 2003b; McCalla, 2000).

Further, informed by user-centered design principles, Preece formulated the community-centered development (CCD) framework to guide practitioners in the field (Preece, 2000). CCD provides a blueprint for building a DCoP. The framework encourages designers to: (1) assess members’ interests, (2) identify community norms and appropriate technology, (3) involve stakeholders in prototype design and testing, (4) correct for poor usability, and (5) foster community network building and identity. Literature informed by this approach draws attention to the interaction between human and technology dimensions in setting the context for the development and sustainability of DCoPs.

CCD integrates a sociotechnical perspective and pays attention to HCI. On the human dimension side, attention has been drawn to understanding participants’ goals, motivations, and perceptions of the learning environment (Daniel et al., 2003b); trust (Preece, 2002); and culture and learning needs (Daniel et al., 2004a). On the technology side, issues include privacy and security, usability, scalability, and authenticity (Daniel et al., 2003a; Preece, 2000).

The attention paid by a sociotechnical approach to HCI makes this framework particularly well

### Table 1. Features of online learning communities and distributed communities of practice (adapted from Daniel et al., 2003b)

<table>
<thead>
<tr>
<th><strong>Formal: Online Learning Communities (OLCs)</strong></th>
<th><strong>Informal: Distributed Communities of Practice (DCoPs)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Membership is explicit and identities are generally known</td>
<td>• Membership may or may not be made explicit</td>
</tr>
<tr>
<td>• Participation is often required</td>
<td>• Participation is mainly voluntary</td>
</tr>
<tr>
<td>• High degree of individual awareness (who is who, who is where)</td>
<td>• Low degree of individual awareness</td>
</tr>
<tr>
<td>• Explicit set of social protocols for interaction</td>
<td>• Implicit and implied set of social protocols for interactions</td>
</tr>
<tr>
<td>• Formal learning goals</td>
<td>• Informal learning goals</td>
</tr>
<tr>
<td>• Possibly diverse backgrounds</td>
<td>• Common subject matter</td>
</tr>
<tr>
<td>• Low shared understanding of domain</td>
<td>• High shared understanding of domain</td>
</tr>
<tr>
<td>• Loose sense of identity</td>
<td>• Strong sense of identity</td>
</tr>
<tr>
<td>• Strict distribution of responsibilities</td>
<td>• No formal distribution of responsibilities</td>
</tr>
<tr>
<td>• Easily disbanded once established</td>
<td>• Less easily disbanded once established</td>
</tr>
<tr>
<td>• Low level of trust</td>
<td>• Reasonable level of trust</td>
</tr>
<tr>
<td>• Lifespan determined by extent in which goals are achieved</td>
<td>• Lifespan determined by the instrumental/expressive value the community provides to its members</td>
</tr>
<tr>
<td>• Pre-planned enterprise and fixed goals</td>
<td>• A joint enterprise as understood and continually renegotiated by its members</td>
</tr>
</tbody>
</table>
suited to understanding the development and sustainability of DCoPs. In particular, the relevance of a sociotechnical approach to the evolution of the GKN project results from the attention to, and monitoring of, feedback loops to inform design and subsequent operation. For example, a sociotechnical approach cautions against a “build it and wait till they come” approach, and favors a co-design process that enables potential users to define their goals and areas of concerns. Joint construction can be regarded as fostering a shared identity and building networks necessary for the development of trust and effective ICT-mediated interaction.

**OUR CURRENT RESEARCH**

The GKN project was launched to address a perceived need to span geography and cross-organizational boundaries to enhance the scholarship on, and the practice of, governance and its role in advancing international development. The underlying challenge of praxis is not unique to this particular subject area. A consultation document issued by the Social Science and Humanities Research Council of Canada, for example, re-stated the networking challenge for advancing collaboration and innovation in the humanities and the social sciences in the following terms:

“Canada is a will against geography. It has a relatively small population, mostly scattered across more than 5,000 kilometres. It has no centres equivalent to Paris or London that naturally draw the best minds and greatest talents…to meet and interact on a regular basis. It does not have the numerous institutions…the Americans have to move people and ideas around. The net result…is that it is hard for people to know each other well, to trust each other and to work together over time and distance.” (SSHRC, 2004)

With the emergence of ICTs, these obstacles to the exchange of information and collaboration were no longer permanent fixtures, though they have tended to endure.

**Research Approach to the Design of User-Centered Online Learning Communities**

We began our effort to overcome these obstacles through a participatory design approach (PDA). Key to PDA is an iterative process that seeks to address users’ needs and promotes their involvement in project development (Schuler & Namioka, 1993). A PDA, also known as a cooperative design approach, shares numerous similarities with Preece’s (2000) community-centered approach.

The first step identified potential technologies capable of spanning geography and nurturing collaboration in a DCoP. Working on the human dimension, the project team created a profile of key stakeholders of 200 individuals from academia, government, and the non- and for-profit sectors. This list represented our target population for the survey of potential users’ views on knowledge sharing in the field and interest in participating in the development of a DCoP.

The users’ assessment was divided into three sections:

- an assessment of existing communication/networking mechanisms among potential community users,
- an assessment of the level of awareness of work undertaken by users and their affiliated organizations, and
- users’ perceived value of a DCoP and what services would contribute to its potential value.

The goal of the users’ assessment was to identify a target group’s interests, perceived
knowledge gaps, thematic content, and potential design models for the proposed GKN portal.

Following the analysis of the assessment, we identified design features that matched identified services together with appropriate technological requirements. We further contacted those who had completed the survey by telephone for a follow-up interview. The goal of the interview was to elicit further information regarding individuals' preferences for content and portal design. These steps also served the equally important objective of engaging potential community participants. In addition, we were able to gauge the reaction to the objectives of the GKN project and method of development and implementation. In addition, the telephone follow-up was an opportunity to initiate informal connections among various individuals working in the same area of research.

RESULTS AND DISCUSSION

The target population for the survey was close to 200 organizations identified as working in the field of international development and governance. The response rate to the survey was 25%. Of those responding, 38% were university based, 23% were from provincial and federal government institutions, 30% were from non-governmental and research organizations, and 9% were from private consulting firms. The respondents were distributed across Canada: 45% from western Canada, 53% from central Canada, and only 2% from the eastern part of the country. These figures reflect the geographical and sectoral diversity of our sample. Four out of five respondents were interested in applied research and technical assistance in this area, and a similar proportion were interested in influencing, contributing, or participating in the policy-making process. In addition, over 80% of respondents revealed that it is important for them to keep current on new developments in research and practice. Depending on their organizational affiliation, 50% to 80% of the respondents were interested in building collaborative partnerships for research and technical assistance

We also asked respondents what kind of research (applied vs. basic research) they were interested in, and if they were willing to share a range of potential outputs with potential GKN members. The majority (90%) responded that they were interested in applied research. They were also willing to contribute to, and participate in, policymaking processes. Participants identified the potential for the GKN to support their interest in keeping abreast of current research and practice in their fields. In terms of collaboration, a large number of the respondents viewed the GKN as a potential mechanism to facilitate information exchange and knowledge sharing among members. These findings were encouraging for, as Lave and Wenger (1991) suggest, CoP development when individuals realize the potential to benefit by sharing knowledge, insights, and experiences with each other, and enhance their practices and performances.

Survey data and follow-up interviews revealed low levels of awareness of contemporary research and practice in the field. At the same time informants commented on the specialized nature of their work and the limited number of organizations active in the field, they also reported that they were largely unaware of contemporary contributions to knowledge and action that their counterparts have made. Though establishing a benchmark of awareness is problematic, our results indicated a considerable lack of awareness among researchers and practitioners working on governance and international development in Canada. The majority of the participants described current knowledge on governance and development as fragmented, and said that there was a serious lack of awareness among people working on similar issues across provinces and between organizations. Similarly, it was observed that a considerable amount of publicly funded research, reports, and policy
documents are not exchanged in a systematic manner. Respondents identified the potential of a GKN initiative to facilitate relations among public, private, non-governmental organizations and academia.

Though overall results revealed that information sharing and knowledge awareness were fragmented, there was a pattern to the responses. First, organizations within a sector were more knowledgeable of current work undertaken by their counterparts within the same sector than organizations in different sectors. Second, there were marked differences in the level of awareness among counterparts within provinces compared to those operating outside their provinces. Although there was a high utilization of information and communication technologies as means to exchange information and data, they were not used systematically to break down the information barriers across organizations and across geographic jurisdictions.

Consistent with previous findings (Wenger, McDermott, & Snyder, 2000), geographic distance is considered an obstacle to knowledge sharing and utilization, even by those who are active users of ICTs. Moving from geographic to language barriers, several respondents underscored the importance of Canada’s two official languages as a potential barrier. Language is critical to any community, since it is deemed as a part of a community identity: identity fosters collaboration and shared understanding within a community (McCalla, 2000).

Turning to services, the following list identifies the top four stakeholder recommendations:

- Design a DCoP to facilitate information exchange and knowledge sharing.
- Provide a platform for sharing lessons, experiences, and best practices.
- Identify and nurture collaboration among government, research community, academia, NGOs, and development practitioners.
- Build linkages and partnerships with other international research communities to advance policy and practice.

Following the analysis of the data and feedback to respondents, we identified and profiled different technologies capable of supporting a DCoP that would perform to stakeholder expectations. Once the technological elements were identified, feedback was sought again from participants on the relevance of these models. This feedback was integrated in the prototype development of the GKN portal, which is currently in its formative stages. As the GKN project moved from a needs assessment to co-development with interested partners, human and technology interaction issues are gaining more importance.

At present, the GKN team has implemented a beta version of the system, while at the same time pursuing research into social and technical means to nurture and support an evolving community. Currently, we are experimenting with the use of blended strategies of face-to-face workshops and videoconferencing as additional avenues to encourage integration of human and technology factors. We are also developing an evaluation plan to assess the importance of the factors identified earlier to developing and sustaining the GKN project. In the following section, we describe the dimensions of HCI that have the potential to affect the viability and robustness of the GKN project.

**EMERGENT HUMAN AND TECHNOLOGY ISSUES**

There are multiple factors affecting the emergence and sustainability of a DCoP. Drawing from the GKN experience and insights from the sociotechnical approach outlined previously, we maintain that the following set of factors are important to HCI. Their influence and relative importance to
the emergence and sustainability of a DCoP is introduced briefly in the following:

- **Didactics:** Learning is a shared experience, and by extension DCoPs are learning communities. Some OLCs have explicit learning goals (e.g., formal OLCs created around WebCT courses), while others have more implicit goals of sharing ideas, practices, and knowledge (e.g., DCoPs among corporate-oriented/professional communities). The technology must therefore enable learning, and perceptions of learning feedback would likely affect participation.

- **Trust:** Stakeholder surveys revealed that a key attraction of the proposed GKN online community would be the ability to share and retrieve archived data that was not widely available. The creation of this shared resource would depend on the willingness of these stakeholders to contribute their data. Their decision to share data would likely be influenced by their trust in others in the community as well as the environment in which they interact, for instance, questions such as: How would community members treat my data? Would my research be reproduced without my permission or quoted out of context? Creating generalized trust within a DCoP is difficult to “engineer”, but likely a pre-requisite condition for the sharing and accumulation of data.

- **Privacy and Security:** Privacy and security tools address individual perceptions of safety in the community. In an environment where a person feels their privacy threatened, declining participation is anticipated. In this regard, computational tools that protect the privacy and security of individuals must be provided.

- **Scalability and Authenticity:** Scalability expresses the ability of a system to accommodate multiple users, and authenticity refers to the ability of a system to protect individuals in a community from outsiders. A DCoP must encourage entrants and their participation. This dimension is critical to the growth of the DCoP, whereas authenticity appears more important to sustainability. For example, an open system that does not protect users (e.g., from hackers) is susceptible to negative feedback and eventual decline of member participation.

- **Sociability:** Sociability relates to the protocols in use for communication and interaction in the community (Preece, 2000). These protocols may be imposed in the first instance, but will likely shift in response to community dynamics. Sociability is of particular importance to “constructed” online communities that do not inherit protocols in use, as would temporal communities that have migrated to an ICT-mediated environment. This dimension is likely critical to the sustainability of a DCoP, as protocols in use will need to reflect members’ preferences and practices. As new protocols emerge, technology must accommodate such changes.

- **Usability:** Our research indicated that interest in the GKN initiative centered on the promise of instrumental outcomes (e.g., access to information, new insights, and expanded contacts). Here, technology and human interaction are clearly linked, as relevant content is dependent on member input and its ease of retrieval is dependent on technology. User-centered interface design and continuous involvement of users are critical to both the emergence and sustainability of the GKN project.

- **Culture:** An explicit objective of the GKN project was to bridge organizational and linguistic boundaries. As organizational theory suggests that organizations inculcate and perpetuate cultures that may promote or discourage inter-organizational infor-
information sharing and/or collaboration. Once organizational or individual participation is present (a human, not a technical issue), we are uncertain of how technology may shape or accommodate different culture(s). Though others suggest that the viability of DCoPs depends on the development of a shared culture, our project is not sufficiently far advanced to comment on this hypothesis.

**Awareness:** The ability of ICT tools to provide awareness among its members is predicted to have a powerful impact on members’ interactions in the community. More specifically, awareness (e.g., awareness about who is who, and who does and knows what) can have a significant positive feedback that would in turn promote participation and contribute to sustainability.

These elements highlighted exert different forces on technology and human interaction. For reasons stated, we anticipate that each will have a bearing on the emergence and sustainability of the GKN initiative and DCoP more generally.

**DISCUSSION**

The sociotechnical approach to the development of a DCoP suggests that human and technical factors are interlinked and they co-determine the emergence, evolution, growth, and sustainability of DCoPs. For practitioners involved in designing or developing a DCoP, the variables outlined previously will likely provide a useful starting point for guiding implementation and identifying key relationships. For researchers, our preliminary exploration of these relationships creates a number of hypotheses for future investigation. As these relationships have a bearing on both practice and research, we intend to track these relationships through user evaluations and internal monitoring. We anticipate that these findings will work toward a framework for comparative research on factors affecting the emergence and sustainability of a DCoP.

By way of conclusion, we offer the following general UCD principles for designing and sustaining online learning communities based on the sociotechnical approach.

**Design Principles**

- Assessing needs of actual or potential users/learners.
- Identifying the gap between what is and what needs to be.
- Understanding users and usage contexts.
- Profiling learning styles.
- Benchmarking existing community models.
- Identifying existing technological tools.
- Maintaining an iterative design and development processes that keep users/learners informed.
- Providing appropriate tools to support and mediate learning, social interaction and facilitate a sense of togetherness.
- Exploring navigation tools to enable members to gather information about others and have access to community interactions traces of activities.

**Didactic Principles**

- Nurturing open and informal discourse as members interact to satisfy their own personal and community learning needs.
- Encouraging learners to become active users and contributors of content.
- Supporting different learning styles.
- Encouraging participation and discourse around central themes, ideas, or purposes.
- Guiding participants throughout the interaction process, and providing them with clear directions to attainment of learning goals.
- Understanding unique individual learning needs differences, and encouraging partici-
User-Centered Design Principles for Online Learning Communities

pants to construct their own meaning based on unique individual experiences.

Sociability Principles

- Establishing a clear set of social protocols for interactions.
- Encouraging informal interaction and an environment conducive to learner/user interaction so that members have opportunities to test the trustworthiness of others.
- Supporting shared objectives—which creates a rationale for belonging to the community.
- Maintaining relevant content and context for interaction throughout the lifespan of the community.
- Encouraging ongoing active dialogue among members.
- Maintaining different forms of awareness (who is who, who knows what, who knows who knows what, etc.) in the community to lubricate effective interaction.

ACKNOWLEDGMENTS

The research reported in this chapter has been supported financially by the Policy Branch of the Canadian International Development Agency (CIDA), the Social Sciences and Humanities Research Council of Canada (SSHRC), and the International Center for Governance and Development at the University of Saskatchewan.

REFERENCES


Riedl, M. O. (2001, January 14-17). A computational model and classification framework for
social navigation. Proceedings of the International Conference on Intelligent User Interfaces 2001 (pp. 137-144), Santa Fe, NM.


ENDNOTE


This work was previously published in User-Centered Design of Online Learning Communities, edited by N. Lambropoulos and P. Zaphiris, pp. 54-70, copyright 2007 by Information Science Publishing (an imprint of IGI Global).
INTRODUCTION

The contribution of context information to content management is of great importance. The increase of storage capacity in mobile devices gives users the possibility to maintain large amounts of content to their phones. As a result, this amount of content is increasing at a high rate. Users are able to store a huge variety of content such as contacts, text messages, ring tones, logos, calendar events, and textual notes. Furthermore, the development of novel applications has created new types of content, which include images, videos, MMS (multi-media messaging), e-mail, music, play lists, audio clips, bookmarks, news and weather, chat, niche information services, travel and entertainment information, driving instructions, banking, and shopping (Schilit & Theimer, 1994; Schilit, Adams, & Want, 1994; Brown, 1996; Brown, Bovey, & Chen, 1997).

The fact that users should be able to store the content on their mobile phone and find the content they need without much effort results in the requirement of managing the content by organizing and annotating it. The purpose of information management is to aid users by offering a safe and easy way of retrieving the relevant content automatically, to minimize their effort and maximize their benefit (Sorvari et al., 2004).

The increasing amount of stored content in mobile devices and the limitations of physical mobile phone user interfaces introduce a usability challenge in content management. The physical mobile
phone user interface will not change considerably. The physical display sizes will not increase since in the mobile devices the display already covers a large part of the surface area. Text input speed will not change much, as keyboard-based text input methods have been the most efficient way to reduce slowness. While information is necessary for many applications, the human brain is limited in terms of how much information it can process at one time. The problem of information management is more complex in mobile environments (Campbell & Tarasewich, 2004).

One way to reduce information overload and enhance content management is through the use of context metadata. Context metadata is information that describes the context in which a content item was created or received and can be used to aid users in searching, retrieving, and organizing the relevant content automatically. Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and the applications themselves (Dey, 2001). Some types of context are the physical context, such as time, location, and date; the social context, such as social group, friends, work, and home; and the mental context, which includes users’ activities and feelings (Ryan, Pascoe, & Morse, 1997; Dey, Abowd, & Wood, 1998; Lucas, 2001).

By organizing and annotating the content, we develop a new way of managing it, while content management features are created to face efficiently the usability challenge. Context metadata helps the user find the content he needs by enabling single and multi-criteria searches (e.g., find photos taken in Paris last year), example-based searches (e.g., find all the video clips recorded in the same location as the selected video clip), and automatic content organization for efficient browsing (e.g., location-based content view, where the content is arranged hierarchically based on the content capture location and information about the hierarchical relationships of different locations).

**DATE, TIME, LOCATION, AND PROXIMITY**

While context can be characterized by a large number of different types of attributes, the contribution of context attributes to content management is of great importance. We focus on a small number of attributes, which are considered the most important in supporting content management and also have the most practical implementations in real products, such as date, time, location, and proximity (nearby Bluetooth devices). Bluetooth is a short-range wireless technology used to create personal area networks among user mobile devices and with other nearby devices.

The first two attributes, date and time, are the most common in use in a wide range of applications. They are used to organize both digital and analog content, and offer an easy way of searching and retrieving the relevant content automatically. For example, many cameras automatically add the date and time to photographs. Furthermore, the location where content is created is another useful attribute for searching the content (e.g., home, workplace, summer cottage). Mobile devices give users the possibility to create content in many different locations. Users can associate the location with the equivalent content in order to add an attribute to it that will enable them to find it easier. Finally, proximity also plays an important role in content management, as nearby Bluetooth devices can provide information both in social and physical context. While each Bluetooth device can be uniquely identified, information can be provided on nearby people by identifying their mobile phones. An example for physical context is the case of a Bluetooth-based hands-free car kit that can be used to identify that the user is in a car.
USABILITY ISSUES AND PROBLEMS

The expansion of the dimension of context information in order to include location, as well as proximity context, can be of benefit to users while they are able to store, access, and share with others their own location-based information such as videos and photos, and feel the sense of community growing among them (Kasinen, 2003; Cheverist, Smith, Mitchell, Friday, & Davies, 2001). But when it comes to proximity to be included in context information, the problem of privacy emerges. It appears that users are willing to accept a loss of privacy when they take into account the benefits of receiving useful information, but they would like to control the release of private information (Ljungstrand, 2001; Ackerman, Darrel, & Weitzner, 2001).

While context metadata is attached to content, when users share content, they have to decide if they share all the metadata with the content or they filter out all or some part of them. The cost for memory and transmission of metadata, as it is textual information, is not an important factor to influence this decision. When the user receives location and proximity information attached to content, he or she may also find out where and with whom the creator of the content was when the content was created. As a result, both the location of the content creator and the location of nearby people are shared along with the content information. If this information is private, the sharing of it could be considered as a privacy violation. This violation may be ’multiplied’ if the first recipient forwards the content and the metadata to other users.

However, users seem to be willing to share context metadata attached to content, as it would be convenient if context metadata were automatically available with the content (so that users do not have to add this information manually). Furthermore, it would be very helpful for the recipient if the received content was annotated with context metadata so that the recipient does not have to annotate it manually and be able to manage the content more easily. For example, in the case of image and video content, the filtering of context metadata such as location and people could be useless, since these same items appearing in the image or video can be identified visually from the image content itself.

But what is meaningful information to the end user? It seems that users want meaningful information, but they are not willing to put too much effort in creating it, unless this information is expected to be very useful. In the case of location, it would be difficult for users to type the name of the place and other attributes manually, since it would require their time and effort. Thus it would be important if meaningful context metadata, which include the required information, are automatically generated.

Proximity information also needs to be meaningful. In this way, meaningfulness is important when attaching information on nearby devices in the form of metadata. If the globally unique Bluetooth device address and the real name of the owner of the device could be connected, this functionality would give meaningful information to the user.

It is hard to determine which information is useful, while what is useful information in one situation might be totally useless in another. For example, when looking at photo albums, what is thought to be useful information varies a lot. When one is looking at family pictures taken recently, it is needless to write down the names of the people, since they were well known and discernable. But it is different looking at family pictures taken many years ago: the same people may not be that easily recognizable.

It appears that useful information depends on a user’s location, what the information is used for, and in which time span. In order to create meaningful information, users need to put much
Content Personalization for Mobile Interfaces

effort into getting the data, organizing it, and annotating it with context metadata. Ways to minimize their effort and maximize their benefit should be developed.

CONCLUSION

The increasing amount of stored content in mobile devices and the limitations of physical mobile phone user interfaces introduce a usability challenge in content management. The efficient management of large amounts of data requires developing new ways of managing content. Stored data are used by applications which should express information in a sensible way, and offer users a simple and intuitive way of organizing, searching, and grouping this information. Inadequate design of user interface results in poor usability and makes an otherwise good application useless. Therefore, it is necessary to design and build context-aware applications.

Issues of usefulness and meaningfulness in utilizing context metadata need to be further investigated. Usefulness depends on the type of metadata. As far as location and proximity are concerned, it appears that the more time has passed since the recording of the data, the more accurate the information needs to be. Furthermore, in the case of location information, the closer to one’s home or familiar places the data refers to, the more detailed the information needs to be. A main usability challenge is the creation of meaningful context metadata automatically, without users having to add this information manually. There exist many ways for automatic recording of information about a user’s context, but the generated information is not always meaningful.

Another field that requires further research is privacy. It seems that users are willing to accept a loss of privacy, provided that the information they receive is useful and they have control over the release of private information. Content management provides users with a safe, easy-to-use, and automated way of organizing and managing their mobile content, as well as retrieving useful information efficiently.

REFERENCES


**KEY TERMS**

**Bluetooth:** A short-range wireless technology used to create personal area networks among user devices and with other nearby devices.

**Content Management:** Ways of organizing and annotating content in order to retrieve and search it more efficiently.

**Content:** Any information that can be used to characterize the situation of an entity.

**Context Metadata:** Information that describes the context in which a content item was created or received.

**Entity:** A person, place, or object that is considered relevant to the interaction between a user and an application, including the user and the applications themselves.

**Location:** The place where content is created by the user.

**Usability:** The effectiveness, efficiency, and satisfaction with which users can achieve tasks in the environment of mobile devices.
Chapter 2.36
Speech–Centric Multimodal User Interface Design in Mobile Technology

Dong Yu
Microsoft Research, USA

Li Deng
Microsoft Research, USA

ABSTRACT
Multimodal user interface (MUI) allows users to interact with a computer system through multiple human-computer communication channels or modalities. Users have the freedom to choose one or more modalities at the same time. MUI is especially important in mobile devices due to the limited display and keyboard size. In this chapter, we provide a survey of the MUI design in mobile technology with a speech-centric view based on our research and experience in this area (e.g., MapPointS and MiPad). In the context of several carefully chosen case studies, we discuss the main issues related to the speech-centric MUI in mobile devices, current solutions, and future directions.

INRODUCTION
In recent years, we have seen steady growth in the adoption of mobile devices in people’s daily lives as these devices become smaller, cheaper, more powerful, and more energy-efficient. However, mobile devices inevitably have a small display area, a tiny keyboard, a stylus, a low speed (usually less than 400 million instructions per second) central processing unit (CPU), and a small amount (usually less than 64MB) of dynamic random-access memory. Added to these limitations is the fact that mobile devices are often used in many different environments, such as dark and/or noisy surroundings, private offices, and meeting rooms. On these devices, the traditional graphical user interface (GUI)-centric design becomes far less effective than desired. More efficient and easy-
to-use user interfaces are in urgent need. The **multimodal user interface** (MUI), which allows users to interact with a computer system through multiple channels such as speech, pen, display, and keyboard, is a promising user interface in mobile devices.

Multimodal interaction is widely observed in human-human communications where senses such as sight, sound, touch, smell, and taste are used. The research on multimodal human-computer interaction, however, became active only after Bolt (1980) proposed his original concept of “Put That There.” Since then, a great amount of research has been carried out in this area (Bregler, Manke, Hild, & Waibel 1993; Codella, Jalili, Koved, Lewis, Ling, Lipscomb, et al., 1992; Cohen, Dalrymple, Moran, Pereira, Sullivan, Gargan, et al., 1989; Cohen, Johnston, McGee, Oviatt, Pittman, Smith, et al., 1997; Deng & Yu, 2005; Fukumoto, Suenga, & Mase, 1994; Hsu, Mahajan, & Acero, 2005; Huang, Acero, Chelba, Deng, Droppo, Duchene, et al., 2001; Neal & Shapiro, 1991; Pavlovic, Berry, & Huang, 1997; Pavlovic & Huang, 1998; Vo, Houghton, Yang, Bub, Meier, Waibel, et al., 1995; Vo & Wood, 1996; Wang, 1995). Importantly, the body of this research work pointed out that MUIs can support flexible, efficient, and powerful human-computer interaction.

With an MUI, users can communicate with a system through many different input devices such as keyboard, stylus, and microphone, and output devices such as graphical display and speakers. MUI is superior to any single modality where users can communicate with a system through only one channel. Note that using an MUI does not mean users need to communicate with the system always through multiple communication channels simultaneously. Instead, it means that users have freedom to choose one or several modalities when communicating with the system, and they can switch modalities at any time without interrupting the interaction. These characteristics make the MUI easier to learn and use, and is preferred by users in many applications that we will describe later in this chapter.

MUI is especially effective and important in mobile devices for several reasons. First, each modality has its strengths and weaknesses. For this reason, single modality does not permit the user to interact with the system effectively across all tasks and environments. For example, speech UI provides a hands-free, eyes-free, and efficient way for users to input descriptive information or to issue commands. This is very valuable when in motion or in natural field settings. Nevertheless, the performance of speech UI decreases dramatically under noisy conditions. In addition, speech UI is not suitable when privacy and social condition (e.g., in a meeting) is a concern. Pen input, on the other hand, allows users to interact with the system silently, and is acceptable in public settings and under extreme noise (Gong, 1995; Holzman, 1999). Pen input is also the preferred way for entering digits, gestures, abbreviations, symbols, signatures, and graphic content (Oviatt & Olsen, 1994; Suhm, 1998). However, it is impossible for the user to use pen input if he/she is handicapped or under “temporary disability” (e.g., when driving). MUI, on the other hand, allows users to shift between modalities as environmental conditions change (Holzman, 1999), and hence, can cover a wider range of changing environments than single-modal user interfaces.

Second, different modalities can compensate for each other’s limitations and thus provide users with more desirable experience (Deng & Yu, 2005; Oviatt, Bernard, & Levow, 1999; Oviatt & vanGent, 1996; Suhm, 1998). For example, the accuracy of a resource-constrained, midsized vocabulary speech recognizer is low given the current speech technology. However, if the speech recognizer is used together with a predictive T9 (text on 9 keys) keyboard, users can greatly increase the text input throughput compared with using the speech modality or T9 keyboard alone (Hsu et al., 2005). The gain is obtained from the mutual disambiguation effect, where each error-
prone modality provides partial information to aid in the interpretation of other modalities. Another reason for the improved user experience is users’ active error avoidance, where users tend to select the input modality that they judge to be less error prone for a particular task and environment (Oviatt & vanGent, 1996), and tend to switch modalities to recover from system errors (Oviatt et al., 1999). Mutual compensation is very important for mobile devices because the ability of every single modality in the devices is extremely limited (e.g., a limited display and keyboard size, and limited speech recognition accuracy).

Despite the importance of MUI in mobile devices, designing effective MUIs is far from trivial. Many MUIs in mobile devices are speech centric, where speech is the central and main modality. In this chapter, we will focus on main issues on the design of effective speech centric MUIs in mobile devices based on our research and experience in developing MapPointS (Deng & Yu, 2005) and MiPad (Deng, Wang, Acero, Hon, Droppo, Boulis, et al., 2002; Huang, Acero, Chelba, Deng, Droppo, Duchene, et al., 2001). In Section 2, we describe a generic MUI architecture in mobile setting that consists of various recognizers for different input modalities, semantic parsers, a discourse manager, and a response manager. In Section 3, we discuss special considerations related to speech modality. In particular, we discuss the approaches to overcoming resource limitations on mobile devices, noise robust speech front-ends, noise robust modality switching interfaces, and context-aware language model. In section 4, we introduce the issues related to robust natural language understanding including construction of robust grammars. We discuss the problem of modality fusion, including modality-neutral semantic representation, unification approach, and modality integration, in Section 5. We discuss possible future directions and conclude this chapter in Section 6.

A GENERIC MUI ARCHITECTURE

The ultimate goal of an MUI is to fulfill the needs and requirements of the users. This principle is one of many emphasized in user-centered design (Gould & Lewis, 1985, Norman & Draper, 1986). According to the user-centered design principle, the acceptability of an MUI can be judged using three main attributes (Dybkjaer & Bernsen, 2001; Hone & Graham, 2001; Nielsen, 1993): effectiveness, efficiency, and learnability. The effectiveness assesses whether users can complete the tasks and achieve the goals with the predefined degree of perceived accuracy. It is usually measured on the targeted user population, over a specified range of tasks and environments. The efficiency judges how much effort (cognitive demand, fatigue, stress, frustration, discomfort, and so on) and resources (time) are needed for users to perform specific tasks. It is usually measured with the total time (including time for error corrections) taken to complete a task. The learnability measures whether users can easily discover the system’s functionality and quickly learn to use the system.

Figure 1 depicts a typical speech-centric MUI architecture that is aimed to achieve a high level of effectiveness, efficiency, and learnability. As shown in the figure, users can communicate with the system through speech, keyboard, and other modalities such as pen and camera. Modality fusion usually is the center of an MUI system. There are two typical ways of fusing information from different input modalities, namely, early fusion and late fusion. With the early fusion, signals are integrated at the feature level and hence, the recognition process in one modality would affect that in another modality (Bregler et al., 1993, Pavlovic et al., 1997; Pavlovic & Huang, 1998; Vo et al., 1995). Early fusion is suitable for highly coupled modalities such as speech and lip movements (Rubin, Vatikiotis-Bateson, & Benoit, 1998; Stork & Hennecke, 1995). However, early fusion
can greatly increase the modeling complexity and computational intensity due to its nature of intermodality influence in the recognition phase. With the late fusion, information is integrated at the semantic level. The benefit of late fusion is its isolation of input modalities from the rest of the system. In other words, individual recognizers trained using unimodal data can be directly plugged into the system without affecting the rest of the system. This feature makes the late fusion easier to scale up to more modalities in the future than the early fusion. The architecture shown in Figure 1 utilizes the late fusion approach that has been widely adopted, for example, by a variety of systems including Put-That-There (Bolt, 1980), MapPointS (Deng & Yu, 2005), MiPad (Huang et al., 2001), ShopTalk (Cohen, et al., 1989), QuickSet (Cohen, Johnston, McGee, Oviatt, Pittman, Smith, et al., 1997), CUBRICON (Neal & Shapiro, 1991), Virtual World (Codella, Jalili, Koved, Lewis, Ling, Lipscomb, et al., 1992), Finger-Pointer (Fukumoto et al., 1994), VisualMan (Wang, 1995), and Jeanie (Vo & Wood, 1996).

In the late-fusion approach depicted in Figure 1, the input signals received by the system are first processed by semantic parsers associated with the corresponding modality into the surface semantics representation. Note that although each modality has its own semantic parser, the resulting surface semantics are represented in a common semantic representation and is thus independent of the modality. The surface semantics from all the input modalities are then fused by the discourse manager component into the discourse semantics representation (more discussions on this issue in Section 4). In order to generate discourse semantics, the discourse manager uses the semantic modal and interacts with the context manager to utilize and update such information as dialog context, domain knowledge, user’s information, and user’s usage history. The updated context information can be used to adapt the language model, which can improve speech recognition accuracy and enhance the quality of semantic parsers for the next user-computer interaction.

The discourse semantics, which is the output of the discourse manager, is then fed into the response manager to communicate back to the user. The response manager synthesizes the proper responses, based on the discourse semantics and the capabilities of the user interface, and plays the response back to the user. In this process, behavior model provides rules to carry out the required actions. The combination of discourse manager and response manager is usually referred to as the dialog manager.

Note that the components shown in Figure 1 may reside on the mobile devices, or distributed on other servers in real implementations. In addition, many MUI systems use an agent-based software solution in which a facility or hub is

Figure 1. A typical speech-centric MUI architecture and its components
used to pass information to and from different components (or agents) (Kumar & Cohen, 2000; Schwartz, 1993).

Many best practices and design principles have been developed for the speech-centric MUI design in the past decades (Becker, 2001; Dybkjaer & Bernsen, 2001; Ravden & Johnson, 1989; Reeves, Lai, J., Larson, J.A., Oviatt, S., Balaji, T.S., Buisine, et al. 2004), which we summarize next.

First, the system should explicitly inform the user about its state through appropriate feedback within a reasonable amount of time, so as to avoid state errors, that is, the user’s perceived state is different from the system’s perceived state. The feedback can be in different modalities, but must be clear and accurate. If speech feedback is used, recorded speech is usually preferred over the synthesized speech, due to its higher degree of naturalness. Note that the recorded speech usually takes a larger amount of resources than the synthesized speech. Since the memory and storage available in mobile devices is very limited, designers should strike a balance between the use of synthesized speech and of recorded speech. The system should follow real-world conventions, and use the words, phrases, and concepts that are familiar to the users. The system should also ensure that the output modalities be well synchronized temporally. For example, the spoken directions should be synchronized with the map display.

Second, the system should provide sufficient flexibility so that users can select the modalities that are best for the task under the specific environments. For example, the user should be able to switch to a nonspeech modality when inputting sensitive information such as personal identification numbers and passwords. A good MUI design should also allow users to exit from an unwanted state via commands that are global to the system, instead of having to go through an extended dialog. The system should provide enough information (e.g., through prompts) to guide novice users to use the system, yet at the same time allow barge-ins and accelerators for the expert users to reduce the overall task completion time.

Third, the system should be designed to allow easy correction of errors. For example, the system should provide context sensitive, concise, and effective help. Other approaches include integrating complementary modalities to improve overall robustness during multimodal fusion; allowing users to select a less error-prone modality for a given lexical content, permitting users to switch to a different modality when error happens; and incorporating modalities capable of conveying rich semantic information.

Fourth, the system’s behavior should be consistent internally and with users’ previous experiences. For example, a similar dialog flow should be followed and the same terms should be used to fulfill the same task. Users should not have to wonder whether the same words and actions have different meaning under different context.

Fifth, the system should not present more information than necessary. For example, dialogues should not contain irrelevant or rarely needed information, and the prompts should be concise.

While the best practices summarized are common to all speech-centric MUIs, some special attention needs to be paid to speech modality and multimodality fusion due to the great variations of mobile device usage environments. We address these special considerations next.

SPECIAL CONSIDERATIONS FOR SPEECH MODALITY

There are two main challenges for the use of speech modality on mobile devices. First, the resources on mobile devices, in particular, CPU speed, memory, and communication bandwidth, are very limited. Second, speech recognition accuracy degrades substantially in realistic noisy environments, where there are abrupt changes in
noise, or variable phase-in phase-out sources of noise as the user moves. For example, the recognition accuracy may drop 30-50% inside a vehicle and cafeteria from that in a quiet environment (Das, Bakis, Nadas, Nahamoo, & Picheny, 1993; Lockwood & Boudy, 1992). Since the mobile devices will be used in these real-field settings without a close-talk microphone, robustness to acoustic environment, that is, immunity to noise and channel distortion, is one of the most important aspects to consider when designing speech-centric MUIs on mobile devices. Speech recognition accuracy and robustness can usually be improved with a noise-robust speech front-end, a noise-robust modality-switching interface, and a context aware language model.

Resource Constrained Speech Recognition

Speech recognition on mobile devices is typically carried out with two options: the distributed recognition (Deng et al., 2002) where the recognition happens at a remote server (Figure 2) and the local recognition (Deligne, Dharanipragada, Gopinath, Maison, Olsen, & Printz, 2002; Varga, Aalburg, Andrassy, Astrov, Bauer, Beaugeant, et al., 2002) where the recognition is carried out completely on the mobile device. The distributed recognition can take advantage of the power of the remote server to achieve a fast and accurate recognition, while the local recognition can eliminate the requirement of the device to have a fast data connection.

In the distributed architecture, the main consideration is the latency required to send data to and from the server. The latency is typically determined by the communication bandwidth and the amount of data sent. To reduce the latency, a typical approach is to use a standard codec on the device to transmit the speech to the server where the coded speech is subsequently decompressed and recognized (as depicted in Figure 3). However, since speech recognizers only need some features of the speech signal (e.g., Mel-cepstrum), an alternative approach is to put the speech front end on the mobile device and transmit only speech features to the server (Deng et al. 2002), as shown in Figure 4. Transmitting speech features can further save bandwidth because the size of the features is typically much less than that of the compressed audio signals.

Besides the advantage of using the computing power at the server to improve speech recognition accuracy, there are other benefits of using server-side recognition. One such benefit is its better maintainability compared to the local recognition approach because updating software on the server is much easier and more cost effective than updating software on millions of mobile devices. It, however, does require the recognizer on the server to be front end or codecagnostic in order to materialize this benefit. In other words, the recognizer should make no assumptions on the structure and processing of the front end (Deng et al., 2002). Another benefit of using distributed recognition is the possibility for the server to personalize the acoustic model, language model, and understanding model all at the server, saving the precious CPU and memory on mobile devices.
In the past, distributed recognition is unquestionably the dominant approach due to the low CPU speed and small amount of memory available on the mobile devices. Nowadays, although the CPU speed and memory size are increasing dramatically, distributed recognition is still the prevailing approach over local recognition due to the advantages discussed previously.

The major issue of the local recognition architecture is the low recognition speed and accuracy due to the slow CPU speed and low memory available on mobile devices. Speech recognizers running on mobile devices need to be specially designed (Deligne et al., 2002, Li, Malkin, & Bilmes, 2006; Varga, Aalburg, Andrassy, Astrov, Bauer, Beaugeant, 2002) to fit the
requirement since speech recognizers designed for the desktop or telephony systems cannot be directly deployed to mobile devices. The greatest benefit of using the local recognition approach is its independency of the network connection and the server and hence, can be used everywhere under any conditions. Given the consistent improvement of the CPU speed and memory on the mobile device hardware, in the future, the local recognition approach is expected to become more and more popular for simple tasks such as name dialing and media playing.

**Noise Robust Speech Front End**

Noise robustness is one of the most important requirements for speech-centric MUI on mobile devices. It has attracted substantial attention in the past several years. Many algorithms have been proposed to deal with nonstationary noises. A popular one is an advanced feature extraction algorithm (jointly developed by Motorola Labs, France Telecom and Alcatel) that was selected in February of 2002 as a standard in distributed speech recognition by the European telecommunications standards institute. The algorithm defines the extraction and compression of the features from speech that is performed on a local, terminal device, for example, a mobile phone. These features are then sent over a data link to a remote “back-end processor” that recognizes the words spoken. The major components of this algorithm are noise reduction, waveform processing, cepstrum calculation, blind equalization, and voice-activity detection. The noise reduction component makes use of two-stage Wiener filtering (Macho, Mauuary, Noé, Cheng, Ealey, Jouvet, et al., 2002).

The stereo-based piecewise linear compensation for environments (SPLICE), which has been used in the MiPad system (Deng et al., 2002), is another effective algorithm for noise robust speech feature extraction. SPLICE is a cepstrum enhancement algorithm dealing with additive noise, channel distortion, or a combination of the two. It is a dynamic, frame-based, bias-removal algorithm with no explicit assumptions made on the nature of the noise model. In SPLICE, the noise characteristics are embedded in the piecewise linear mapping between the “stereo” clean and distorted speech cepstral vectors. SPLICE has a potential to handle a wide range of distortions, including nonstationary distortion, joint additive and convolutional distortion, and nonlinear distortion (in time-domain), because SPLICE can accurately estimate the correction vectors without the need for an explicit noise model.

**Modality Switching**

One of the problems in speech recognition under noisy environment is modality switching. If the speech recognition engine is always on, noises and by-talks may be misrecognized as a legitimate user input and hence, can erroneously trigger commands.

A widely used modality switching approach is called “push to talk,” where the user presses a button to turn on the speech recognizer, and releases the button to turn off the recognizer. Another approach is called “tap & talk” (Deng et al., 2002; Huang, Acero, A., Chelba, C., Deng, L., Duchene, D., Goodman, et al., 2000, Huang et al., 2001), where the user provides inputs by tapping the “tap & talk” field and then talking to it. Alternatively, the user can select the tap & talk field by using the roller to navigate and holding it down while speaking. Tap & talk can be considered as a combination of push-to-talk control and indication of where the recognized text should go. Both the push-to-talk and tap & talk avoid the speech detection problem that is critical to the noisy environment under which the mobile devices are typically deployed.

Figure 5 shows an example of the tap & talk interface used in the MiPad (Deng et al., 2002). If the user wants to provide the attendee information for a meeting scheduling task, he/she taps
the “attendees” field in the calendar card. When that happens, the MUI will constrain both the language model and the semantic model based on the information on the potential attendees. This can significantly improve the accuracy and the throughput. Note that tap & talk functions as a user-initiative dialog-state specification. With tap & talk, there is no need for the mobile devices to include any special mechanism to handle spoken dialog focus and digression.

**Context-Aware Language Model**

Here, *context* refers to any information that can be used to characterize the situation related to human-computer interaction. It typically includes the surrounding environment (e.g., location and noise condition), the user (e.g., age and gender, preferences, past interaction experiences, and the interaction history in the current session), and the devices (e.g., remaining battery life, available memory, screen-size, screen-contrast, and speaker volume). Although context-awareness can be beneficial to all components in an MUI, it is especially important for improving speech recognition accuracy under noisy environments.

Context information can be utilized in many different ways in speech modality. One particular approach is to construct the language model based on the context. For example, the tap & talk approach (Deng et al., 2002) customizes the language model depending on the field the user is pointing to, as mentioned in section 3.3. Language model can also be customized, based on the user information and the dialog state. For example, if the system is expecting the recipient information, the language model can include only the names in the global address book. If the user information is also used, the language model can also include user’s contact list and people who have exchanged e-mails with the user in the past. An even more effective language model would weight different names differently, depending on the frequencies the user exchanged e-mail with the person, and the recentness of the interaction (Yu, Wang, Mahajan, Mau, & Acero, 2003). Another example of constructing the language model based on the context and user informa-

---

*Figure 5. An example of the Tap & Talk interface (Deng et al., 2002, © 2002 IEEE)*
tion is described in the speech enabled MapPoint (Deng & Yu, 2005). Without context information, the speech recognizer needs to load all location names and business names in the North America. This is definitely beyond the ability of most state-of-the-art speech recognizers. However, if the user’s location information and/or the interaction history are known, the system can load only the location names and business names around the user’s current location, and weight all the names based on the popularity of the names as well as the user’s interaction history.

A more advanced context-aware language model construction technique is discussed by Wang (2004). This detection-based technique is used in the second generation of the MiPad (Wang, 2004). The basic idea of this approach is to detect the context cues from the user’s partial utterances sequentially, and adjust the language model dynamically for the next part of the utterances. This approach has achieved excellent user experience.

**LANGUAGE UNDERSTANDING**

Good speech recognition accuracy does not always translate to good understanding of users’ intents, as indicated by Wang, Acero, and Chelba (2003). A robust language-understanding model is needed to obtain good user experience for speech-centric MUI applications, especially since speech recognition errors will affect the understanding.

The first issue to address in language understanding is constructing the semantic grammar. Since the importance of each word to the understanding is different, the words need to be treated differently. A typical approach is to introduce a specific type of nonterminals called semantic classes to describe the schema of an application (Wang, 2001; Yu, Ju, Wang, & Acero, 2006). The semantic classes define the concepts embedded in the linguistic structures, which are usually modeled with probabilistic context-free grammars. The advantage of introducing the semantic classes is to make the linguistic realization of semantic concepts independent of the semantic concepts themselves. Once the semantic classes are defined, a robust linguistic grammar can be built using the approaches similar to the one described by Yu, et al. (2006).

The transformation from the recognized text to the semantic representation is usually done using a semantic parser. For example, in MiPad, this transformation is done using a robust chart parser (Wang, 2001). In this parser, “the robustness to ungrammaticality and noise can be attributed to its ability of skipping minimum unparsable segments in the input. The algorithm uses dotted rules, which are standard context free grammar rules in Backus Naur form plus a dot in front of a right-hand-side symbol. The dot separates the symbols that already have matched with the input words from the symbols that are yet to be matched.” (Wang, 2001, pp. 1556) Since the language models used in MiPad are dynamically generated based on the current user information and the tap & talk field, the parser used in MiPad supports dynamic grammars. Given that some part of the user’s utterances is in the free-style form (e.g., the topic of a meeting to be scheduled), they are modeled as dictation grammar rules. Since speech recognition is not perfect, the MiPad robust parser takes into account the N-best list, together with the associated confidence scores returned from the speech recognition engine, and combines the speech recognition score with the parsing score to obtain the best parsing result. More recent progress includes using maximum entropy models to classify the tasks and to disambiguate the meaning of the slots in the recognition result.

**MODALITY FUSION**

One strong advantage of using MUIs is the improved accuracy and throughput through modality integration. There are typically two fusion ap-
proaches: early fusion and late fusion. Given that late fusion has many superior properties over the early one, as discussed in Section 2, it will be the focus of our discussion in this section. There are two tasks in the late fusion: Process and convert the input signals into a common surface semantic representation using the semantic parsers (one specific to each modality), and fuse the surface semantics into discourse semantics using the discourse manager.

**Semantic Representation and Unification**

The semantic fusion operation requires a meaning representation framework that is common among modalities, and a well-defined operation for combining partial meanings.

Many semantic representation formats have been proposed in the past. For example, in Bolt’s (1980) pioneering paper, only very limited modality fusion is required and hence, a simple semantic representation was used. In the past decade, researchers (Cheyer & Julia, 1995; Pavlovic & Huang, 1998; Shaikh, Juth, Medl, Marsic, Kulikowski, & Flanagan, 1997; Vo & Wood, 1996) have converged to using a data structure called typed feature structures (Kay, 1979) to represent meanings. Typed feature structure can be considered as an extended, recursive version of attribute-value-type data structures, where a value can, in turn, be a feature structure. It extends frames (Minsky, 1975) that represent objects and relations as nested sets of attribute/value pairs, by using shared variables to indicate common substructures. A typed feature structure indicates the kind of entity it represents with a type, and the values with an associated collection of feature-value or attribute-value pairs. In the typed feature structure, a value may be nil, a variable, an atom, or another typed-feature structure.

The primary operation on typed feature structure is *unification*. "Typed-feature-structure *unification* is an operation that determines the consistency of two representational structures and, if they are consistent, combines them into a single result." (Oviatt, Cohen, Wu, Vergo, Duncan, Suhm, et al., 2000, online version pp. 21) Unification can combine complementary input from different modalities and rule out contradictory input (Johnston, 1998).

Note that users’ multimodal inputs may involve sequentially integrated or simultaneously delivered signal fragments. In other words, temporal relationships between different input channels are very important. To fuse modalities, we need to first determine whether two input fragments are related. In most of the systems reported, this is achieved by considering all input contents that lie within a predefined time window. To do this, all input fragments need to be time stamped as soon as they are generated to remove the errors due to transit delays.

For example, the speech input “Show me the restaurants around here.” might have a gesture-input accompanying it either “before,” “during,” or “after” the actual utterance, and all these three possibilities should provide the same result. Usually the term “before” represents a timeframe of up to several minutes, “during” represents a timeframe of 4 to 5 seconds, and “after” represents a timeframe of 500ms to 750ms. If these values are too small, many multimodal inputs will be considered as unimodal inputs and will not be integrated. If the values are too large the chances of an old or invalid user input are likely being accepted as part of a valid multimodal input.

To determine whether two input fragments should be treated as parts of a multimodal construction or separate unimodal commands, knowledge gained from a user study is very helpful. For example, it has been shown in Oviatt, DeAngeli, and Kuhn (1997) that users’ written input precedes speech during a sequentially integrated multimodal command. They have also clarified the distribution of typical intermodal lags.
Semantic Fusion with Uncertain Inputs

The challenge of semantic fusion with uncertain inputs is to determine the unified meaning based on multimodal input fragments associated with probabilities. This is especially important for speech-centric MUI because the output of a speech recognizer is never certain. Note that the unification operation on the typed feature structure assumes that all input modalities are certain, and so they cannot be directly applied here. To fuse modalities with uncertainties, a hybrid symbolic/statistical architecture that combines statistical processing techniques with a symbolic unification-based approach is in need. This combined approach involves many factors when fusing the semantics. These factors include recognition accuracy of the individual modalities, the way of combining posterior probabilities, and the prior distribution of multimodal commands.

Note that a multimodal input gives rise to three different types of information overlay: nonoverlayed, overlayed and nonconflicting, and overlayed and conflicting. Nonoverlayed information indicates that the input (unimodal or multimodal) does not have any of the same information represented multiple times. This is the simplest condition. Overlayed and nonconflicting information refers to information segments that may have been represented multiple times without a conflict. The overlayed and conflicting information refers to the case that the information has been provided multiple times and conflicts. There are many approaches to resolving conflicting information in typed feature structure if no uncertainty is involved. The “unification” approach simply returns the value null when a conflict is detected. The “overlay” method returns the first argument when conflicting information is present. However, given that the semantic information from different modalities should not be equally trusted, a better conflicting information resolving approach can be found to handle input signals that may or may not be overlapped in their temporal delivery (Oviatt et al., 1997). Note that overlayed information may arise when inputs are from different modalities (e.g., speech and gesture), or when the same-type modality information occurs multiple times over an extended time frame. Both these two conditions need to be handled.

Conventionally, the probability of the merged feature structures is the cross product of the probabilities of individual feature structures based on the assumption that inputs are statistically independent with each other. In this section, we describe an alternative statistical approach that has been used in QuickSet (Wu, Oviatt, & Cohen, 1999). This approach uses the associative map to reduce the unification pairs and members-teams-committee (MTC) model to refine the multimodal integration process so that different weights are assigned to different modes and different constituents.

**Associative map** defines all semantically meaningful mapping relations that exist between different sets of constituents for each multimodal command. In its simplest form, it can be considered as a simple process of table lookup. For example, if an MUI consists of only the speech modality and the pen modality, we can build a two-dimensional table. If two inputs from different modalities can be fused, the value at the corresponding cell is 1; otherwise, the value is 0. The purpose of the associative map is to rule out considerations of those feature structures that cannot possibly be unified semantically.

**Members-teams-committee** weighs the contributions derived from different modality recognizers based on their empirically-derived relative reliabilities. MTC consists of multiple members, multiple teams, and a committee. “members” are the individual recognizers that provide a diverse spectrum of recognition results (local posterior probabilities). Member recognizers can be on more than one team. Members report their results to their recognizer team leader, which then applies various weighting parameters to their reported
scores. Furthermore, each team can apply a different weighting scheme, and can examine different subsets of data. Finally, the committee weights the results of the various teams, and reports the final recognition results. The parameters at each level of the hierarchy are trained from a labeled corpus.” (Oviatt, et al., 2000, online version, p. 24).

**CONCLUSION AND FUTURE DIRECTIONS**

In this chapter, we discussed the importance of using the MUI in mobile devices, and described the state-of-the-art technologies in designing speech-centric MUI in mobile devices. Specifically, we discussed the noise robustness technologies, the reliable modality switching methods, the context-aware language model, and the robust language-understanding technologies that contribute to the usability of the speech modality. We also described the modality integration technologies that are important to improving the accuracy and throughput of the MUI. Although these technologies have greatly advanced the speech centric MUI design and development in the mobile devices, future research is needed in the following areas.

**Microphone Array Processing**

Noise robustness is still a challenging research area for speech-centric MUIs. Although many single-microphone noise robustness technologies (e.g., Deng, et al., 2002; Macho, et al. 2002) have been proposed to improve speech recognition accuracy under noisy environments, the progress so far is still limited. Given the continuous decrease in the hardware price, using microphone array on mobile devices is a trend to combat noisy acoustic conditions and to further decrease speech recognition errors. Microphone array algorithms, which take advantage of the received signal differences between microphones, can achieve noise suppression of 10-15 db effectively (Tashev & Malvar, 2005). Future research is needed for more efficient and effective algorithms using low-cost, low-quality microphone arrays that may be equipped in speech-centric mobile devices.

**Error Handling Techniques**

Fragile error handling continues to be a top interface problem for speech-centric MUI (Karat, Halverson, Horn, & Karat, 1999; Rhyne & Wolf, 1993; Roe & Wilpon, 1994). A great amount of research work needs to be done in developing graceful error-handling strategies in speech-centric MUI. First, new statistical methods need to be developed to reduce errors through mutual disambiguation between modalities. Second, new dialog strategies (e.g., mixed initiative) need to be developed to allow easy correction of the errors. Third, the system needs to be able to adapt to different environments and challenging contexts to reduce errors. Fourth, better robust speech recognition technologies need to be developed to increase the speech recognition accuracy under a wide range of environments.

**Adaptive Multimodal Architectures**

In most current MUI systems, their behaviors are predesigned by the developers. The system does not automatically learn to improve the performance as users use the system. Given that mobile devices are usually used by a single user, it is very important to develop adaptive MUI architectures.

For example, Oviatt (1999) showed that any given user’s habitual integration pattern (simultaneous vs. sequential) is apparent at the beginning of their system interaction. When the user uses the system, the interaction pattern remains the same. An adaptive MUI system that can distinguish and utilize these patterns to improve the modality fusion could potentially achieve greater recognition accuracy and interactive speed. Another
example is for the system to gradually change the behavior (e.g., automatically predict the user’s next action) when the user changes from a novice to an experienced user.

Future research in this area would include what and when to adapt, as well as how (e.g., through reinforcement learning) to adapt MUI systems so that their robustness can be enhanced.

**Mixed Initiative Multimodal Dialog**

Most current speech-centric MUI systems are user initiative, where the user controls the dialog flow (for example, through push to talk). A user-initiative system can be modeled as a set of asynchronous event handlers. In a more advanced system, the system should also actively interact with the user to ask for missing information (which is called mixed initiative). For example, if the user wants to search for the phone number of a business using a mobile device and he/she forgets to mention the city and state information, the dialog system should automatically ask the user for that information through the multimodal output devices.

Future research should address the design and development of consistent and efficient conversational interaction strategies that can be used by different multimodal systems. Multimodal dialogue systems should be developed within a statistical framework (Horvitz, 1999) that permits probabilistic reasoning about the task, the context, and typical user intentions.

**REFERENCES**


speech processing in MiPad’s multimodal user interface. *IEEE Transactions on Speech and Audio Processing*, 10(8), 605-619.


KEY TERMS

Modality: A communication channel between human and computer, such as vision, speech, keyboard, pen, and touch.

Modality Fusion: A process of combining information from different input modalities in a principled way. Typical fusion approaches include early fusion, in which signals are integrated at the feature level, and late fusion, in which information is integrated at the semantic level.

Multimodal User Interface: A user interface with which users can choose to interact with a system through one of the supported modalities, or multiple modalities simultaneously, based on the usage environment or preference. Multimodal user interface can increase the usability because the strength of one modality often compensates for the weaknesses of another.

Push to Talk: A method of modality switching where a momentary button is used to activate and deactivate the speech recognition engine.

Speech-Centric Multimodal User Interface: A multimodal user interface where speech is the central and primary interaction modality.

Typed feature Structure: An extended, recursive version of attribute-value type data structures, where a value can, in turn, be a feature structure. It indicates the kind of entity it represents with a type, and the values with an associated collection of feature-value or attribute-value pairs. In the typed feature structure, a value may be nil, a variable, an atom, or another typed feature structure.

User-Centered Design: A design philosophy and process in which great attention is given to the needs, expectations, and limitations of the end user of a human-computer interface at each stage of the design process. In the user-centered design process, designers not only analyze and foresee how users are likely to use an interface, but also test their assumptions with actual users under real usage scenario.
Chapter 2.37
Kinetic User Interfaces:
Physical Embodied Interaction with Mobile Ubiquitous Computing Systems

Vincenzo Pallotta
University of Fribourg, Switzerland

Pascal Bruegger
University of Fribourg, Switzerland

Béat Hirsbrunner
University of Fribourg, Switzerland

ABSTRACT

This chapter presents a conceptual framework for an emerging type of user interfaces for mobile ubiquitous computing systems, and focuses in particular on the interaction through motion of people and objects in physical space. We introduce the notion of Kinetic User Interface as a unifying framework and a middleware for the design of pervasive interfaces, in which motion is considered as the primary input modality.

INTRODUCTION

Internet and mobile computing technology is changing the way users access information and interact with computers and media. Personal Computing in its original form is fading and shifting towards the ubiquitous (or pervasive) computing paradigm (Want et al., 2002). Ubiquitous Computing systems are made up of several interconnected heterogeneous computational devices with different degrees of mobility and computing power. All of these devices and appliances are embedded in everyday objects, scattered in space, capable of sensing the environment and of communicating with each other, and carried or exchanged by people. Therefore, we are facing a new ecology of computing systems that poses new issues in their integration and usability. Human-computer interfaces that were designed for desktop personal computers must be re-conceived for this new scenario. Due to the different capabilities...
of mobile and embedded devices, the pervasive computing infrastructure, and the nature of their expected usage, it is apparent that new types of user interfaces are needed in order to unleash the usability of new generation distributed computing applications (see (Rukzio, 2006) for a classification of mobile devices interfaces). Additionally, the concept of user interface itself seems to be no longer adequate to cope with ubiquitous computing systems. Rather, it is the concept of interaction and user experience that will take over (Beaudouin-Lafon, 2004).

**Ubiquitous Computing**

Ubiquitous Computing (henceforth Ubicomp) is an emerging research sub-area of Distributed Systems whose main focus is studying how heterogeneous, networked computing devices can be embedded in objects of daily use in order to enable new applicative scenarios and user experiences. Mark Weiser (1991; 1993; 1994) introduced the term Ubiquitous Computing in the ‘90s as a new way to understand computer technology and to lay the foundations of an expected and necessary computing paradigm revolution. Weiser’s vision has been adopted and interpreted by a great number of researchers, among whom we consider relevant for our goals the works of (Abowd & Mynatt, 2000; Abowd et al., 2002; Banavar & Bernstein, 2004; Bellotti et al., 2002; Greenfield, 2006; Norman, 1999; Want et al., 2002). We summarize the Ubicomp vision in four fundamental points that motivate our effort of providing a new conceptual framework for Ubicomp user interfaces:

1. Today’s computer (e.g., the personal computer) will disappear, and the computing power will fade inside the network infrastructure, as it is already the case to some extent with existing web-services.
2. Computing will be extremely distributed and heterogeneous. This will result from the interconnection of several computing devices, each specialized in specific tasks and scattered in the physical environment (ranging from embedded devices to high-performance servers).
3. Computer interfaces will no longer capture the full attention of users. Rather, computer applications will run in “background” most of the time, accomplishing “routinized” operations, and they will try to gain user’s attention only when strictly required.
4. Computer interfaces will be unobtrusive and based on new emerging interaction models obtained by direct interaction with physical objects and with the whole environment.

Ubiquitous Computing is often equated to (or better, confused with) nomadic computing (Kleinrock, 1997). Nomadic computing is a form of computing environment that offers its users access to data or information from any device and network while they are in state of motion. In nomadic computing, the use of portable devices (such as laptops and handheld computers) in conjunction with mobile communications technologies enables users to access the Internet and data on their home or work computers from anywhere in the world. Mobile connectivity certainly does play an important role in Ubicomp, but it is not the only one. We consider of central importance the user’s mobility intended as the user’s ability of moving objects and themselves in the physical space. In fact, in using Ubicomp systems, users are no longer forced to sit in front of a desktop computer and to operate it with mice, keyboards and local input/output devices. Users will interact through actions performed on everyday objects that surround them. As pointed again by Weiser:

> [ubiquitous computing] is different from PDAs, dynabooks, or information at your fingertips. It is invisible, everywhere computing that does not live on a personal device of any sort, but is in the woodwork everywhere.
Kinetic User Interfaces

User Interfaces for Ubicomp Systems

Human-computer interaction (HCI) is a very large domain of research, which includes all related aspects of interaction with digital appliances. Although HCI has been mostly focused on graphical user interfaces (GUIs), we consider here those aspects that pertain to how users will interact with computers in the 21st century (Winograd, 2000), that is, by direct manipulation of objects in their physical environment. Paul Dourish (2001) defined this paradigm as embodied interaction. According to his definition, meaningful embodied interaction with a digital system can be obtained only if an alignment is maintained between the physical and the digital world. In other terms, “bits” and “atoms” must live together in peace (Negroponte, 1995). As recently pointed out in (Sparacino, 2005, p. 2):

[…] computation and sensing are moving from computers and devices into the environment itself. The space around us is instrumented with sensors and displays, and this tends to reflect a widespread need to blend together the information space with our physical space.

Embodied interaction is thus aimed at exploring new interaction patterns where people are exploring how to move the interface “off the screen” and into the real world (Shafer et al., 2001). For instance, in Tangible User Interfaces (TUIs) (Ulmer and Ishii, 2000; Holmquist et al., 2004), tangible interaction is intended to replace desktop GUI’s interaction and elements with operations on physical objects. The motion of objects in the physical space determines the execution of actions, such as item selection (by means of what in TUI are called “phicons” i.e., physical icons), service requests, database updates, and so forth. Rekimoto (1997) proposed the Pick&Drop pattern, an extension of the Drag&Drop pattern, to move items across computers. In his work on graspable user interfaces, Fitzmaurice (2000) proposed to extend the interaction with classical GUI by means of physical objects (such as LEGO bricks) over an augmented desktop surface. Tangible and graspable user interfaces are undoubtedly a great achievement in HCI. However, they are strongly biased by GUIs interfaces; nearly no new types of interaction induced by the nature of the physical space and objects have been proposed other than replications of those available on ordinary desktop GUIs.

The development of the ideas introduced by TUIs, combined with the ubiquity of information over the Internet, has led to a new concept of what counts as an Ubicomp system. Moving from the assumption that only information appliances could constitute a ubiquitous system, the whole material world of things and people now can be made computational and connected. The movement of the “Internet of Things” (ITU, 2005; Bleeker, 2006) is aimed at promoting the idea that any object can have a virtual identity in the Internet realm, as long as the object can embed a unique identifier, which corresponds to an IP address. This is now possible because of the larger 128 bits IP address space offered by the new IPv6 protocol (i.e., \(3.4 \times 10^{38}\) addresses). According to (Greenfield, 2006), the type of infrastructure that enables Ubiquitous Computing is already technically feasible and it will soon scale up to a level that will make it possible to safely connect a huge number of small heterogeneous devices, possibly embedded in everyday objects.

Augmented Reality (AR) (Mackay, 1998) and Wearable Interfaces (Barfield & Caudell, 2001) are emerging technologies that support embodied interaction for Ubicomp systems. We believe that Augmented Reality focuses more on how feedback is provided, whereas Wearable Interfaces focuses on the types of devices that can support embodied interaction.
**Unobtrusive Interfaces**

When HCI intersects Ubicomp, many assumptions that were made when designing interaction for ordinary computing devices are no longer valid. In Ubicomp, computers exist in different forms and only in a minimal portion as ordinary desktop computers (i.e., where interaction is performed through screens, keyboards, mice). As pointed out by Weiser and other promoters of Ubicomp, interacting with a ubiquitous system should be realized through an *unobtrusive interface*, more precisely, an interface that does not capture the full attention of the user, who can still use the system to perform the foreground tasks (Nardi, 1996). In contrast, an *obtrusive interface* is one that requires an unjustified cognitive effort to be operated, thus interfering with the normal usage of the system. Weiser & Seely Brown (1996) call this setting “Calm Technology” in order to stress the importance of adapting the computers and their interfaces to human pace, rather that the other way around. In this vision, computers should follow users in their daily activity and be ready to provide information or assistance on demand. Moreover, they should not require much attention from the user by asking information that can be autonomously obtained from the actual usage context. They must be “aware” of the context and be able to adapt their behaviour and interfaces to different usage situations. In other words, ubiquitous computers must be *smart* and *adaptive*.

**Kinetic-Awareness**

Context-awareness is considered as the most important issue in Ubicomp (Baldauf et al., 2006; Dourish, 2004; Hong et al., 2005). Specifically, location-awareness is considered a key component of context in designing user interfaces for mobile systems. Location-awareness has been always treated as a sub-case of context-awareness, and motion as a form of context change. Location change is taken as a context change for adapting the application’s behaviour, rather than as an explicit intentional act within the application’s interface.

We believe that location changes occurring over time can represent more than just context-change. It can be considered as input modality. This does not mean that location or motion context has to be neglected. Motion input and location context can be used together in the same way as in handheld GUIs, where interaction with the mobile devices can be contextualized through location or motion. For instance, it does make sense to consider a different interpretation for an object’s motion that occurs in a different location, or to interpret the GUI input differently when the mobile device is moving with different kinetic properties. To clarify this aspect let us consider two examples.

The first example is about two possible situations where a paraglide is flying i) over a lake, or ii) over the ground. Motion input is treated accordingly to the actual situation; when flying over the lake some manoeuvres are considered risky, while they are considered safe if flying over the ground (landing is possible everywhere).

The second example is a situation where implicit interaction with GUI is enabled if the handheld device is detected moving with a certain speed. For instance, suppose that a dialogue box asking for a confirmation pops up on the handheld’s GUI. If the user is moving, at a speed of more than 5 Km/h, then after a timeout, the interface assumes that the default choice has been (implicitly) selected. If the user is moving at lower speed or is still, the interface will wait for input without the timeout.

According to Dix et al. (2000), *space* and *location* define a new design space for interactive mobile systems. Mobile devices have increasing capacity in providing location and motion information as part of their usage context. They also acknowledged that accounting for motion as an input modality for location-aware mobile Ubicomp systems opens potential research op-
portunities. In fact, they propose and instantiate a taxonomy of the degrees of mobility with available technologies and applications. They categorize degrees of mobility along three dimensions level of mobility (i.e., the physical bound of the device to the environment), level of independence (i.e., the relation of the used location-aware device with other devices or to the environment), level of co-operativeness (i.e., the extent to which the device is bound to a particular individual or group). Some existing mobile devices and applications are then categorized according to this three-dimensional taxonomy. Some difficult or even impossible cases, such as a fixed-pervasive-personal, that might be instantiated by a fully isolated active cell (e.g., a prison cell with motion sensors). The combination that is relevant for us is that of pervasive-mobile-personal/group devices. Such devices might be location/motion-aware objects that can be used to interact with Ubicomp systems.

Systems combining location and motion awareness, henceforth, will be referred to as kinetic-aware systems. Our analysis of how kinetic information is taken into account in existing Ubicomp systems revealed that there are two main views on the role of kinetic information in user interfaces:

1. In the first view, location and motion are used as a component of the usage context that is exploited by applications running on the mobile information computing devices for adapting their behaviour. In this view, there is no physical interaction with the place (and its contained artefacts) where the users are using their own devices. The environment is not “sensing” the presence and the physical action of the user. Moreover, neither the system nor the environment is supposed to handle spatio-temporally located events.

2. In the second view, location change and motion are considered as primary input modalities reflecting the user’s goals and intentions while using Ubicomp applications. That is, users can intentionally perform explicit and implicit actions through physical motion. These actions are recognized and contextualized in the place where they occur, by possibly affecting the state of co-located devices and remote systems. Interaction through motion with physical space becomes the main focus, rather than simply contextualizing applications based on ordinary user interfaces running on mobile devices.

While the majority of research on location-aware Ubicomp systems focuses on the first view on kinetic awareness, we focus instead on the second view, which can be achieved by recognizing users’ goals and intentions from various properties of motion of objects in the physical space through what we call the Kinetic User Interface (KUI).

Kinetic awareness can be seen as part of a more general paradigm, defined in (Dix et al., 2004) as context-aware computing. This emerging paradigm poses several challenges. One of the biggest recognized difficulties for this type of system is interpreting human activity. We do not propose algorithms for automatic recognition of human activities (which is a machine learning task). Instead, we propose a framework in which activities can be decomposed into smaller and easier-to-detect patterns. These patterns are instantiated by acquiring input from sensors that are linked to software components representing physical objects. We believe that providing a level of abstraction for activities will make easier the design and the implementation of Kinetic User Interfaces for Ubicomp systems.

**Related Works and Technologies**

Location-aware services are nowadays available to mobile Internet users. With the advent of Web2.0, a great deal of applied research and development recently has been devoted to embedding Internet
technology into everyday life mobile devices, ranging from pure entertainment to critical applications such as healthcare, national security, military (Cáceres et al., 2006). The rationale behind these efforts is to provide the mobile Internet users with great flexibility in authoring, publishing, and retrieving information, as well as in accessing services that are relevant in a given situation and place. A remarkable example of Web2.0 mobile application for multimedia information retrieval is SocialLight3, which allows the tagging of geographical location with multimedia tags (e.g., shadow-tags). The roaming users can “geo-tag” a place either by using a mobile phone when they are physically present there, or by attaching the tag on the SocialLight Web page with a GoogleMap mash-up. Information can further be retrieved from the Web site through the mashed-up map or by querying the system by using a mobile, GPS-enabled Internet device. Geo-tagging is also the focus of the Mobile Multimedia Metadata project of the Garage Cinema Lab (Davis et al., 2004), whose main purpose is to cooperatively annotate geo-located pictures. (Ashbrook et al., 2006) push this concept further and propose a roadmap for future research with a new type of scenario based on location-awareness and motion tracking, in which users can capture media while moving, and share their experience by making the captured media stream available to Internet users in real-time.

Location-awareness in Ubicomp has been the main focus of several projects since 1990. Among the projects that heavily rely on location context are the Aware Home (Kidd et al., 1999) at GeorgiaTech, the GUIDE project at Lancaster University (Chervest et al., 2000), the AURA project at Carnegie Mellon University (Sousa & Garlan, 2002), the GAIA’s Active Space project at University of Illinois Urbana Champaign (Román et al., 2002), the Interactive Maps at ETHZ (Norrie, 2005), and the Global Smart Places project (Meyer et al., 2006). It might seem that this is a widely explored research area, but a closer analysis reveals that all of these projects deal with a very basic form of motion-awareness. Location change is taken as a user’s context change, which is used for dynamically adapting the application’s behaviour. They all consider a user’s location as an additional parameter of an explicit service request. Then, the service’s output is delivered on the handheld device carried by the user or on a nearby display.

The Cyberguide project (Abowd et al., 1996) is one among the first attempts in taking a user’s motion into account (see also (Schilit, 1995)). A tourist equipped with indoor (IR beacons) and outdoor (GPS) localization devices can automatically receive contextual relevant information on a PDA while moving, and feed a trip journal.

In the EasyLiving project at Microsoft (Brumitt et al., 2000), a geometric model of the physical space is used in order to enable physical embodied interaction by representing the physical relationships between entities in the world. Unfortunately, this model has a limited scope since it is adapted to room scale, and it only considers current spatial relationships between objects, while ignoring their motion within the tracked space.

The Intelligent Workspace project at MIT (Koile et al., 2003) is a system that records the user’s location history and learns the so-called activity zones. Activity zones are portions of the physical space where the user is doing specific activities and which are repeatedly observed by the system. Once an activity zone is learned, the settlement of user in it will automatically trigger a number of pre-defined services that support the observed activity. In other words, the system reacts to context change and, in particular, to the user’s location change.

However, none of the above-mentioned systems explicitly recognize motion as the primary input modality and as a mean of performing a purposeful (explicit or implicit) action within the application’s interaction space. The following systems make a further step towards a more explicit notion of kinetic-awareness.
In the Sentient Computing project at AT&T Cambridge Research labs (Addlesee et al., 2001), motion tracking is an essential feature for interacting with the system. The system is built around the ActiveBat infrastructure for tracking location and motion using ultrawide-band sensors (now commercialized by Ubisense®). A few applications for the ActiveBat have been proposed, such as “FollowMe” that allows users to move their input-output environments over several devices scattered in the environment (e.g., phone call forwarding, virtual desktop displacement), or the “Virtual Mouse” that allows users carrying the ActiveBat device to use it as a mouse over a wall display. A relevant aspect of this project for us is the adopted context modelling techniques. The application-level and the user-level (mental) model of the environment are kept aligned by the system (Harter et al., 2002). In other words, when users perform actions, they update their mental representation of the resulting state of the environment by directly observing it. This representation might not be consistent with the information that applications have gathered through sensors. The system must take care of checking and possibly restabilising the lost alignment.

The Sonic City project (Gaye et al., 2003) exploits motion in the urban landscape as a way for interactively creating a musical experience. The user motion is tracked, as well as the current position over the city map. Motion and location contexts are combined with other contexts obtained through wearable sensors in order to influence the composition of music content in real-time. Users of the Sonic City interface can hear the result of musical composition during their walking activity.

Other applications of the motion context are pervasive games and races. Games that involve location tracking (like trails in (Spence et al., 2005)) are suited for exploiting the motion context. For instance, in CatchBob (Nova et al., 2006), a multi-player game developed at EPFL, the players’ motion is tracked and their paths are made visible to other team members on a digital map. The goal of the game is to cooperatively perform a task by looking at the motion of the other team members on a handheld display and by communicating direction suggestions in real-time.

Overview of the Chapter

In this chapter, we propose a unifying framework for the design and implementation of Kinetic User Interfaces for Ubicomp systems by (i) defining a set of fundamental concepts and (ii) by presenting a middleware that enables the use of motion in physical spaces as the primary interaction modality.

The remainder of this chapter is structured as follows. In Section 2, we provide some backgrounds and intuitions about the distinguishing features of KUIs compared to other types of user interfaces. We also present a few motivating applicative KUI-based scenarios. In Section 3, we detail the main concepts of KUI and we present the structure of its middleware architecture. Conclusions are finally discussed in Section 4.

Kinetic User Interfaces

The term “Kinetic” is derived from the Greek *kinetikos*, which means “moving of, relating to or resulting from motion (the action or process of moving).” In physics, kinetic theory explains the physical properties of matter in terms of the movement of its constituent parts; kinetic energy refers to energy, which a body possesses by virtue of being in motion. Kinetic abilities of humans are of no question. People move and change their current spatial location all the time and in a mostly unconscious way. Humans are also capable of “modulating” motion in several ways, by keeping or varying their speed, by following different trajectories or patterns (e.g., dancing), or by executing various types of motion in parallel (e.g., gesturing while walking).
At different scales and contexts, motion (or absence of motion) can be recognized as a purposeful action. For instance, if a tourist stops long enough in front of a statue, it might be reasonable to assume that he or she is observing the monument. What if the statue (or the environment) would be smart enough to provide the tourist with relevant information about its author or style, or, even smarter, to figure out that the tourist has already stopped at the statue, to avoid repeating the old information unless the tourist explicitly requests it?

A familiar everyday situation occurs when items are passed through a bar code reader in a grocery store counter. Their motion (the passage in a given place) is recognized as a purchase transaction. Another typical KUI situation takes place when somebody, possibly unconsciously, performs dangerous actions such as moving into dangerous areas. In these cases, a monitoring system could alert the user by signalling the potential danger.

**KUI vs. GUI**

Motion in the physical space is such a common and pervasive phenomenon that we hardly recognize its status as an interaction modality with a computer. While it is apparent that motion plays an essential role in WIMP\(^5\) Graphical User Interfaces (GUI), the virtual and limited nature of GUI's space, the “desktop,” seems not to afford the full bodily motion interaction (Beaudoin-Lafo, 2000). For instance, it does make little or no sense to talk about speed or acceleration of the pointer. Sometimes, however, these properties are taken into account by specific types of PC applications like games or flight simulators.

We introduce the concept of *Kinetic User Interface* (KUI) as a way of endorsing Weiser’s Ubiquitous Computing vision (Weiser, 1993) and the Dourish’s Embodied Interaction vision (Dourish, 2001) discussed in the previous section. Accordingly, KUIs are intended to enable a new interaction model for pervasive computing systems in which the motion of objects and users in the physical space are recognized as events and processes to which the system reacts. To make a parallel with ordinary, pointer-controlled Graphical User Interfaces (GUIs), moving the pointer on the display and clicking on a graphical item is recognized by the system as an intentional act, which usually triggers a system’s reaction on the software representation of the domain object associated to the selected graphical item and, possibly, the execution of an action on the domain object specified in the application currently running on the computer. Similar to “hovering” the pointer over a desktop in GUIs, in KUIs users can trigger input events for the computing environment by moving themselves and by displacing tracked objects. Users can exploit the physical space by executing actions/operations on physical objects, such as moving, grabbing, touching, juxtaposing, whose effects are reflected in the application objects. For instance, following a path or executing a pre-defined motion pattern can be viewed as similar to “mouse gestures” in ordinary desktop GUIs and can consequently trigger reactions by the gesture-enabled application.

KUIs are not limited to single-user interfaces and do not impose a unique locus of interaction. Hence, it enables richer interactions than GUIs and it is better suited to ubiquitous and mobile computing environments. Motion as an input modality can be used alone or in combination with other input modalities available to the user for interaction with the system, which are directly afforded by other mobile devices carried by the users and by fixed input devices located in the interaction space (e.g., ordinary point and click, or speech recognition).

**Feedback Management**

As in GUIs, an important issue in KUIs is feedback management. Due to the different nature of physical space with respect to GUI’s synthetic
Kinetic User Interfaces

In the physical space, feedback cannot be provided in the same way as for GUIs. Since one of the goals of KUI is to help build unobtrusive interfaces, we give back to users only the minimal amount of information required to inform them that their interaction with the physical space has been successfully recognized. In turn, the system should avoid interfering with the user’s current activity if the effects of the recognized action have only a peripheral importance to the current foreground task (i.e., if the effects impact only objects that are not in the current focus of attention of the user). Moreover, since the physical space already allows for the direct manipulation of real objects, feedback should only inform users about those effects produced in the computing space to (virtual) domain objects.

Although users might not always be aware of what effects are caused by their motion and the motion of tracked objects, they will be unobtrusively notified when their motion has been detected and interpreted by the system. Different than GUIs or even in Augmented Reality systems, there will be no need to display a synthetic image of the moving object. The only graphical components of the interface will be those corresponding to additional modalities. For example, we can imagine a scenario where a dialog box is prompted on a mobile device or on an embedded display when the user is detected to walk by or stop at a specific point of interest.

A feedback mechanism of control is also necessary for other reasons, such as privacy; to grant a certain level of protection, users must be somehow aware when their presence and motion is being currently tracked. Consequently, they must always be given the possibility to stop the tracking of the mobile device and to be allowed to use an alternative interaction modality.

KUI Interaction Patterns

Although KUI interaction patterns can be radically different from GUI patterns, some of the most effective GUI patterns, such as Drag&Drop, can be transferred and adapted to KUI interaction with physical space. For instance, in a KUI-enabled SmartHome, the user can “drag” the media being currently played in the living room and “drop” it to the bedroom just by moving a representative localizable object such as the remote controller. It is worth noting that the “Drag&Drop” pattern is provided as an interaction pattern by the KUI middleware and can be activated (and recognized) for specific applications such as the SmartHome control system.

Another useful pattern we include in KUI is continuous tracking. Continuous physical motion is comparable to mouse-gestures in GUIs. KUI-enabled applications are supposed to recognize certain kinetic patterns that might be naturally performed by users during other activities or specific situations.

As an example of the continuous tracking pattern, consider the scenario where the user is driving a car and some of the car’s motion parameters are obtained by embedded sensors such as a GPS tracking system and an accelerometer. The sensors reveal that the car is decelerating in the proximity of a gas station (i.e., a geo-located point of interest already known by the application). This kinetic pattern (deceleration) is detected by the KUI and interpreted by the application as the user’s intention of refuelling at the gas station. This hypothesis might be corroborated by other contextual information from the current car’s sensors (e.g., the fuel level being almost zero). As a result of this behaviour, the system will pro-actively prompt the driver with the current gas prices at the approaching gas station. The application might also perform further contextual inferences and inform the user that keeping the current speed and considering the current fuel level he/she can reach the next gas station that has better gas prices. However, if the system detects that the fuel level is high or the fuel tank is even full, it will not react because it can infer that the
driver stops for other (unknown) reasons (e.g., to take a pause).

This is a clear example of how KUI interaction differs from ordinary location-aware user interfaces. In this scenario, the driver passing by a gas station does not need to explicitly inquire about the gas prices. Current location information is only used to contextualize the query that is triggered as a result of the speed change occurring in the proximity of a given point of interest.

With regard to enabling technologies, continuous motion tracking is already available with current GPS-based car navigation systems, and easily can be integrated with personal mobile devices (e.g., SmartPhones, PDAs) connected to mobile Internet infrastructures (such as UMTS, GPRS, WiMax). With ordinary GPS navigation systems, the user can always check the current location on a graphical map and the proximity to point of interests. With KUI, we extend the possibilities of these systems with an additional level of interactivity and integration with networked services (Pallotta et al., 2006).

KUI-Enabled Scenarios

KUI-enabled applications have a different purpose compared to ordinary location-aware ones. Since motion is used as an input modality, KUI-based applications are expected to provide a higher level of fluidity in interaction and user experience. In those situations in which the interface should not interfere with the foreground user’s activity (which in turn might or might not be a computer-based one), KUI will allow unobtrusive interaction with a computer. We consider here three case studies that have been developed so far and that exemplify the benefits of KUI interaction patterns.

Safety in Air Sports

Jane is flying with her paraglide over the Alps and is trying to reach the other side of a lake she is currently over. Can she do it without any risks?

The UbiGlide flight navigator detects the motion of the paraglide. By interpreting her current activity UbiGlide infers Jane’s intention to cross the lake. UbiGlide then senses the environment, namely, the wind’s force and direction, the lake altitude, the distance between the paraglide and the opposite shore, and finally concludes that the crossing is not possible. Jane is informed immediately about the risk of danger. Later, she is so focused on the flight that she finds herself approaching a no-fly zone (e.g., an airplane landing strip). UbiGlide detects this possibility and alerts her about the danger.

In this scenario the current paraglide motion is not only tracked but also interpreted. An activity report is obtained by composing a number of basic flight movements in order to recognize more complex behaviour patterns. For instance, spiral ascension is made of several turns and altitude changes. Moreover, activities are filtered by other contextual information such as wind speed and direction, air humidity, and so forth.

SmartHome and SmartCity

Julia, Steve, and their daughter Monica live in Geneva. Julia is a busy businesswoman and Steve is a researcher at the university. Monica is in her last year of college. They live in a flat equipped with latest IT technology. They own a networked SmartFridge, which detects when food items go in and out, and automatically generates a list of missing items. The UbiShop system looks at the list generated by the SmartFridge and sends requests to buy the missing items to any family members who pass by a grocery store. Monica is on her way back home after school and passes by a grocery store. A reminder to buy milk is sent on her mobile phone by the UbiShop system. She decides to ignore the reminder since she knows that another grocery store is on her way home. Meanwhile, Steve is also on his way home and near a grocery store. He also receives the reminder and
decides to buy the milk. This purchase causes the deletion of the milk from the shopping list, so that Monica will no longer be bothered. When back home, Steve does not put the milk in the fridge. After a while, the SmartFridge “wonders” why the milk has not yet been put inside, so a request about this item is sent to Steve who had simply forgot the milk in the car.

This illustrates how KUI can contribute to context-aware collaboration in a mixed urban/home environment. Family members’ activities are coordinated according to actual members’ mobility. The system decides to adapt a workflow by interpreting team members’ behaviors in context. Here, the role of the KUI interface is twofold. First, it allows triggering the task assignment when the user is moving into a zone where the task could be accomplished. Second, it detects from the user’s speed whether he/she is likely to be willing to perform the assigned task. For instance, if the user is running, the application could interpret this motion pattern as “being in a hurry” and might decide not to bother the user. Another interesting aspect where KUI plays a role is when the user is expected to perform an action and this action does not occur. This is the case when, in our scenario, Steve forgot to put the milk in the fridge. The application subscribed to a motion event that does not occur within a time interval.

Safe-Critical Work Environments

Bill is a specialized worker in a chemical plant. He typically operates an industrial chemical reactor. He wears a head-mounted display and he is connected to the main control desk through wireless radio communication gears. While normally operating the reactor, suddenly Bill starts running (accelerates) toward the emergency exit. He is not able to alert the control desk about what is happening. The KUI interface detects this motion pattern as abnormal and figures out that the operator is trying to escape from a dangerous situation. The system then opens the doors on the pathway toward the exit of the reactor building and then immediately closes them after the operator is sensed to have passed through them.

In certain industrial settings such as chemical plants, it is important that the operator keeps his hands free in order to be able to do his usual manual work, while at the same time he accesses the automated commands and looks at the supervision information needed to complete the task. The role of KUI is apparent because it provides an additional implicit input modality that might serve to detect sudden instinctive reactions to dangerous situations.

THE KUI MODEL

In this section, we present the main concepts of Kinetic User Interfaces that are implemented as software components in the KUI middleware architecture.

The KUI Conceptual Taxonomy

In KUI, motion is a main (or primary) interaction modality afforded by the physical space to users through the motion of tracked entities. Tracked entities are any objects or autonomous (possibly living) things for which we can provide location and motion information. Tracked entities are represented by KUI components called Kuidgets. Interaction with Kuidgets happens when users affect their motion properties or change spatio-temporal relationships among them (e.g., an object is entering into an area). For instance, when the user is driving a car, the motion properties of its corresponding Kuidget will be continuously updated with its current position, speed, acceleration, and direction.

The term “Kuidget” has been chosen to make the parallel with a GUI’s widgets, that is, soft-
ware components that provide public interfaces for a hardware sensor and whose interaction is implemented in terms of messages and call-backs (Winograd, 2001). Kuidgets are the software counterpart of some real world entities that can be used for interaction in KUIs. A KUI-enabled system is thus able to recognize the current location of Kuidgets, and makes sense of their motion parameters such as path, speed, acceleration, and direction. Location and motion sensors (e.g., GPS or other tracking devices, accelerometers, compasses, altimeters) typically provide three-dimensional location and motion information to Kuidgets. Kuidgets are classified according to four main dimensions: geometry; kinetic properties; degree of autonomy; and type of motion.

From the geometric point of view, Kuidgets can be arbitrary three-dimensional objects. However, it makes sense to distinguish between those objects whose size is not relevant and those for which it matters. Entities of the first type are considered as points while others are considered as geometric shapes. This distinction is application-dependent, because one entity can be considered a point in one application and a shape in others. For instance, a vehicle (a car, a train, a ship, a plane) is a point when considered as a moving object in the space, and it is a space when considered as a container of objects and people. In the KUI model, the same entity plays two roles at the same time and is linked to two distinct Kuidgets.

Kuidgets can be fixed or mobile. Fixed Kuidgets are typically places or landmarks in the physical space, while mobile Kuidgets are physical entities whose location and motion can be observed by tracking the entity or can be provided by their embedded location and motion sensors.

In modelling KUI’s dynamics we adopt the status-event semantics, which means that KUI-based applications should be able to effectively deal both with the status of objects and with the events they generate. Thus, as an underlying model for dealing with Status-Event semantics in KUI, we adopt the Environment and Artefacts theory for multi-agent systems as proposed by (Ricci et al., 2006). Accordingly, we further classify Kuidgets along their degree of autonomy as artefact Kuidgets and agent Kuidgets. Artefact Kuidgets correspond to mobile physical objects that cannot move autonomously (e.g., a mobile phone, a car). The motion properties of an artefact Kuidget can be directly determined and communicated by the artefact itself (e.g., a GPS sensor + a mobile network connection) or observed by another entity (e.g., the infrastructure of the containing object, a nearby artefact, or an agent). The events of moving, grabbing, dragging, and dropping artefact Kuidgets are triggered by the detection of their current kinetic properties in the physical space and by the actions performed through their interfaces for direct manipulation (e.g., pressing a button on the object while moving it). Agent Kuidgets correspond to autonomous moving entities (people, animals, robots). Agent Kuidgets have a higher degree of autonomy, and they can induce motion to other artefact Kuidgets.

At the physical level, there is not much difference between artefacts and agent Kuidgets; they are essentially KUI’s components, and as long as their corresponding physical objects are providing their location and motion information, they are treated equally in KUIs. At a conceptual level, however, they differ because agents can control and operate artefacts and have a higher degree of autonomy. This distinction is particularly useful when KUI-enabled applications need to determine the location and motion of Kuidgets in the absence of up-to-date information. Motion of artefact Kuidgets typically has to be somehow causally linked to agent Kuidgets; sometimes artefacts cause the motion of agents, and sometimes it is the other way around.

Artefact Kuidget typically keeps its last observed (communicated) location if they are not linked to any moving entity (i.e., the law of inertia). Even in cases where no information can be obtained for an artefact Kuidget, its current location can be inferred by default just by knowing that
Kinetic User Interfaces

it is unlinked from a moving Kuidget. Different from artefact Kuidgets, agent Kuidgets have more freedom. When unlinked from any moving artefact Kuidgets, their current location and motion status cannot be safely inferred; they can move without being tracked. Of course, there also might be the case that an artefact is no longer tracked, but this case is considered as an error condition rather than a choice selected by the user. In other words, in KUI agent Kuidgets can decide when they can be tracked, while artefacts cannot.

Two Kuidgets can be logically linked and they can provide location and motion information to each other. For instance, a car equipped with a GPS sensor (an artefact Kuidget) can provide location and motion information to its driver (an agent Kuidget). Conversely, a user carrying a GPS sensor can provide location and motion information to the car Kuidget by setting a logical link between the car and the user who is driving the car. It is important to notice that when the user leaves the car, even if the link is destroyed, the car Kuidget keeps its last location.

It is also our goal to make KUI as general as possible in order to uniformly cope with different geographical scales (e.g., tabletop, room, building, cities) and with different types of location-aware devices (e.g., GPS, RFID, Wireless cell triangulation, ultrasonic, ultra-wideband, infrared). For this purpose, we also distinguish between different types of physical space. Following (Dix et al., 2000), we consider topological and symbolic spaces. In topological spaces, objects are localized by their exact position by means of an absolute coordinate system and through a notion of distance. In symbolic spaces, locations are considered as symbolic elements (e.g., rooms, buildings, cities), and object are localized through spatial relations with other objects (e.g., in the bedroom, near the red table). For topological spaces, the two basic types of references are points and zones, while for symbolic spaces, entities are explicitly connected through symbolic spatial relations such as containment, accessibility, and so forth. Different from (Dix et al., 2000), we do not make any distinction between real and virtual spaces. In our approach, virtual spaces are managed by applications and they do not need to be represented at the KUI level. However, KUI allows the geo-localization of virtual objects; these are entities of the computing space mapped to the geographical space (e.g., a geo-tag or a zone).

The detection of particular spatio-temporal relations between Kuidgets can trigger application-specific KUI events. There are several spatio-temporal relations that can be modelled. We propose here that a basic KUI should provide at least two types of spatio-temporal relations: proximity and containment. For these relations it is important to consider their temporal dimension, namely the start and end time, and the duration of the relation. For instance, if two mobile Kuidgets (e.g., two agents) are observed while moving together along the same path or into the same location, the application will be notified with a “joint motion” event by the KUI manager. Then, the application might make inferences, and as a result, establish an application-dependent relation between the two Kuidgets. For instance, when the two agent Kuidgets are moving together, the application can infer (with the help of other contextual information) that they might be friends. Similarly, when two Kuidgets that are supposed to jointly move cease to do so, an event could be triggered that in certain circumstances could denote an unusual situation. For instance, a car moving somewhere while its owner moves elsewhere else might denote that the car has been stolen.

The last dimension for our classification is the type of motion of mobile Kuidgets. In order to cope with most possible situations, we consider both endogenous and exogenous motion. Endogenous motion occurs when objects move without any change of location. For instance, a rotating object that remains in the same place counts as endogenous motion; another example is the sudden motion sensors embedded in Apple Macintosh laptops that can be exploited as an input modal-
ity in games. Exogenous motion represents the familiar case in which objects are displaced in the space. However, even if an entity is affected by exogenous motion, this does not necessarily mean that the exact change of location is tracked. For instance, if motion is detected by an accelerometer, the interface can use this information to trigger an event. This is the case of a Nintendo Wii™ controller WiiMote that is used to interact with games without detecting the exact change of location of players. More precisely, endogenous motion pertains to the fact that the spatial coordinate system is centred on the moving object, while exogenous motion is referenced to an external coordinate system.

The KUI Middleware

The software architecture we propose for KUI is supposed to be integrated within a larger context-aware middleware for Ubicomp. As in GUI, KUI should be independent from the underlying OS, and should enable rich interaction for context-aware Ubicomp systems. For this reason, we do not commit ourselves to a specific middleware. Rather, we focus on a general “pluggable” software component, which would allow us to make KUI available in an arbitrary context-aware middleware (see Baldauf et al., 2006) for a review). However, KUI can be used as a standalone system if the interface between the KUI manager and motion-aware mobile devices is made through ad-hoc drivers.

The KUI middleware is made of three layers of abstraction (as shown in Figure 1). Below, we provide details about each layer:

- The Observation Layer is the lower level. Its role is to collect kinetic contexts from location and motion aware devices and from other hardware sensors.
- The KUI-Space Layer is an object-oriented environment that contains and manages Kuidgets state and their semantic relationships. Information flow coming from the observation layer is used to update the state of Kuidgets. Location information pertaining to Kuidgets is stored in a suitable data

Figure 1. KUI middleware
Kinetic User Interfaces

structure, the GeoDB (e.g. a GIS\(^1\)), that is also used to store and manage physical space references of fixed and mobile Kuidgets for both topological and symbolic spaces.

- The Activity Layer manages the context history, and aggregates KUI events into higher-level semantic events (i.e., for the detection of specific interaction patterns) that are sent to the applications. In this layer, representations of situations will be constructed by aggregating kinetic information from Kuidgets and other contextual information. Moreover, models of activities will be matched with the spotted situations in order to determine the occurrence of anomalous behaviours.

More specifically, in the KUI software components location and motion information are linked to either fixed Kuidgets or mobile Kuidgets that are localizable by means of tracking hardware.

The KUI Components

The KUI Middleware can be seen as a user interface server that connects client applications to motion-aware input devices. In that sense, our model is similar to the X-Windows toolkit model. KUI provides an API to applications for subscribing aggregated events from the Activity Layer and encapsulates motion-aware devices through the Observation Layer. The main components of the KUI middleware are detailed in Figure 2.

A number of location and motion-aware devices can be connected to any Kuidget for which they cooperate in providing its kinetic information. Kuidgets are identified by universally unique identifiers (UUID), and are linked to entries in

Figure 2. Sketch of the KUI toolkit
the GeoDB that provides both direct localization and indirect localization through explicit relationships with other Kuidgets not linked to directly localizable elements. When the Observation Layer produces fresh kinetic information for a given Kuidget, a callback is sent to KUI-Space manager component, which updates the corresponding geographical information in the GeoDB and manages the relations between active Kuidgets.

The KUI-Space Manager is responsible for the aggregation and filtering of low-level events produced by the Observation Layer. Observations also can be obtained from the hosting context-aware middleware through context widgets (Dey et al., 2001). Context widgets are abstractions for different types of context sensors. They hide the complexity of the sensor communication protocols and offer a hot-plugging mechanism for dynamically adding and removing sensors to the system. As for GUIs, the application does not have to be modified when the pointing system changes, for instance, from mouse to pen. In our case, Kuidgets connected to location and motion Widgets do not know what kind of sensor is actually in use when they get kinetic information.

The Relation Manager receives location updates from the KUI-space and processes them according to the (programmable) relation rules. Relations between Kuidgets are created, deleted or updated and dispatched to the KUI-space manager, which then sends them to the upper Activity Layer through a notification mechanism. When some pre-defined geographical relationship change occurs (e.g., an object enters an active zone), the Relation Manager is responsible of notifying these events to the Activity Layer or directly to the subscribing Kuidgets. These events can be programmed as triggers that check conditions matched against the GeoDB, while aggregate events can be subscribed by objects in the upper Activity Layer. Relations between Kuidgets can be created either by the Kuidget’s internal logic in response to Kuidget events triggered when retrieving information from the GeoDB, or explicitly by applications. The latter case is useful when we logically link agent and artefact Kuidgets together, allowing one of them to inherit motion properties from the other one.

The Activity Layer is responsible of aggregating motion information from one or several Kuidget, as well as dynamic information generated by the Relation Manager. Relations and Kuidgets status are used as the building blocks for the recognition of the previously described kinetic interaction patterns.

Enabling Technology

The KUI middleware can be implemented on top of a context-aware middleware such as the Context Toolkit (Dey et al., 2001) and it can be integrated within any enterprise architecture like J2EE or .NET. Kinetic-aware devices typically will be connected through wireless Internet so that client-server software architecture is apparently justified. Applications exploit localization infrastructures for indoor and outdoor tracking. Indoor localization technologies include RFID antennas, ultrasonic, ultrawide-band, and IR sensors. For outdoor localization and motion tracking, GPS offers the most available tracking solution, which, combined with wireless Internet communication (e.g., GPRS, EDGE or UMTS) is nowadays available on commercial mobile phone and handheld devices. Additionally, we expect to detect others (more local) motion parameters (such as acceleration and direction) by using wearable sensors like accelerometers and digital compasses. For this point, it is crucial for the Observation Layer to be capable of dealing with several location and motion tracking technologies at the same time and of easily associating them with Kuidgets. The accuracy of different localization devices and motion sensors is not considered to be an issue in this discussion that pertains to the conceptual framework for the development of user
Kinetic User Interfaces

interfaces based on kinetic input. We expect, in any case, that motion tracking and wearable sensor technology will rapidly improve, as already has been the case for other enabling technology like, for instance, the diffusion of broadband Internet connection for the Web.

CONCLUSION

In this chapter, we explored the notion of kinetic-awareness in Ubicomp user interfaces by means of the seamless and transparent integration of objects motion detection in the physical space as a primary input modality. Kinetic User Interfaces enable the users of Ubicomp systems to establish an interaction through continuous tracking of kinetic-aware mobile devices at different spatial scales and by the acquisition of kinetic input through motion-aware embedded sensors. KUI interfaces allow the seamless integration of contextual (implicit) and intentional (explicit) interaction through motion. We presented a conceptual framework for KUI interfaces and a middleware as the basis for implementing the KUI component in standard Ubicomp architectures.

Kinetic-awareness in Ubicomp seems to take over simple location-awareness. Motion-based interaction is a complementary notion to context-awareness. It is not just a matter of acting while moving, but acting by moving. Motion is a great source of information that leverages new dimensions of user experience in Ubicomp systems. As noted in (Beaudouin-Lafon, 2004), in post-WIMP user interfaces, it will be necessary to shift towards a more holistic view of user interaction. Users are expected to interact through activities rather than single actions. Moreover, they will try to achieve higher-level goals through activities, rather than to accomplish tasks through planned actions. KUI provides a framework for designing Ubicomp applications with embodied interaction with a special focus on unobtrusiveness and fluidity.

REFERENCES


ENDNOTES

1 http://www.ubiq.com/hypertext/weiser/UbiHome.html
2 http://en.wikipedia.org/wiki/Web_2.0
3 http://www.socialight.com
Kinetic User Interfaces

http://www.ubisense.com

WIMP stands for “Windows, Icons, Menus, Popups”.

The motion detection can be obtained either by the mobile device itself (e.g. a GPS-enabled handheld) or by external device or infrastructure (e.g. a badge tracked by a sensing space).

This interaction pattern is similar to the Teleport application (Addlesee et al., 2001), which allows users wearing ActiveBadges to move their desktop environments from a PC to another.

See (Salkham et al., 2006) for an overview.

We changed the names attributed to these types of space by (Dix et al., 2000). Our “topological” corresponds to their “cartesian” and our “symbolic” corresponds to their “topological”.

http://www.apple.com/

Geographical Information Systems

This work was previously published in Advances in Ubiquitous Computing: Future Paradigms and Directions, edited by S. Mostefaoui, Z. Maamar, and G. Giaglis, pp. 201-228, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
This section presents extensive coverage of specific tools and technologies that humans interact with and react to in their daily lives. These chapters provide an in-depth analysis of devices and tools such as portable music players, mobile phones, and even blogs. Within these rigorously researched chapters, readers are presented with countless examples of the technologies that support and encourage societal development and their resulting impact.
INTRODUCTION

Bluetooth (Bluetooth SIG, 2004) and ZigBee (ZigBee Alliance, 2004) are short-range radio technologies designed for wireless personal area networks (WPANs), where the devices must have low power consumption and require little infrastructure to operate, or none at all. These devices will enable many applications of mobile and pervasive computing. Bluetooth is the IEEE 802.15.1 (2002) standard and focuses on cable replacement for consumer devices and voice applications for medium data rate networks. ZigBee is the IEEE 802.15.4 (2003) standard for low data rate networks for sensors and control devices. The IEEE defines only the physical (PHY) and medium access control (MAC) layers of the standards (Baker, 2005). Both standards have alliances formed by different companies that develop the specifications for the other layers, such as network, link, security, and application. Although designed for different applications, there exists some overlap among these technologies, which are both competitive and complementary. This article makes a comparison of the two standards, addressing the differences, similarities, and coexistence issues. Some research challenges are described, such as quality of service, security, energy-saving methods and protocols for network formation, routing, and scheduling.

BLUETOOTH

Bluetooth originated in 1994 when Ericsson started to develop a technology for cable replacement between mobile phones and accessories. Some years later Ericsson and other companies joined together to form the Bluetooth Special Interest Group (SIG), and in 1998 the specification 1.0 was released. The IEEE published the 802.15.1 standard in 2002, adopting the lower
Enabling Technologies for Pervasive Computing

layers of Bluetooth. The specification Bluetooth 2.0+EDR (Enhanced Data Rate) was released in 2004 (Bluetooth SIG, 2004).

Bluetooth is a low-cost wireless radio technology designed to eliminate wires and cables between mobile and fixed devices over short distances, allowing the formation of ad hoc networks. The core protocols of Bluetooth are the radio, baseband, link manager protocol (LMP), logical link control and adaptation protocol (L2CAP), and service discovery protocol (SDP). The radio specifies details of the air interface, including frequency, modulation scheme, and transmit power. The baseband is responsible for connection establishment, addressing, packet format, timing, and power control. The LMP is used for link setup between devices and link management, while the L2CAP adapts upper-layer protocols to the baseband layer. The SDP is concerned with device information and services offered by Bluetooth devices.

Bluetooth operates on the 2.4 GHz ISM (Industrial, Scientific, and Medical) band employing a frequency-hopping spread spectrum (FHSS) technique. There are 79 hopping frequencies, each having a bandwidth of 1MHz. The transmission rate is up to 1 Mbps in version 1.2 (Bluetooth SIG, 2003) using GFSK (Gaussian frequency shift keying) modulation. In version 2.0+EDR new modes of 2 Mbps and 3 Mbps were introduced. These modes use GFSK modulation for the header and access code of the packets, but employ different modulation for data. The $\pi/4$ differential quadrature phase-shift keying (DQPSK) modulation and 8 differential phase-shift keying (DPSK) modulation are employed in 2 Mbps and 3 Mbps mode, respectively.

The communication channel can support both data (asynchronous) and voice (synchronous) communications. The synchronous voice channels are provided using circuit switching with a slot reservation at fixed intervals. The asynchronous data channels are provided using packet switching utilizing a polling access scheme. The channel is divided in time slots of 625 $\mu$s. A time-division duplex (TDD) scheme is used for full-duplex operation.

Each Bluetooth data packet has three fields: the access code (72 bits), header (54 bits), and payload. The access code is used for synchronization and the header has information such as packet type, flow control, and acknowledgement. Three error correction schemes are defined for Bluetooth. A 1/3 rate FEC (forward error correction) is used for packet header; for data, 2/3 rate FEC and ARQ (automatic retransmission request). The ARQ scheme asks for a retransmission of the packet any time the CRC (cyclic redundancy check) code detects errors. The 2/3 rate FEC is a (15,10) Hamming code used in some packets. The ARQ scheme is not used for synchronous packets such as voice.

Figure 1. Piconet and scatternet
The devices can communicate with each other forming a network, called piconet, with up to eight nodes. Within a piconet, one device is assigned as a master node and the others act as slave nodes. In the case of multiple slaves, the channel (and bandwidth) is shared among all the devices in the piconet. Devices in different piconets can communicate using a structure called scatternet, as shown in Figure 1. A slave in one piconet can participate in another piconet as either a master or slave. In a scatternet, two or more piconets are not synchronized in either time or frequency. Each of the them operates in its own frequency-hopping channel while the bridge nodes in multiple piconets participate at the appropriate time via TDD. The range of Bluetooth devices depends on the class power, ranging from 10 to 100 meters.

ZIGBEE

ZigBee has its origins in 1998, when Motorola started to develop a wireless technology for low-power mesh networking (Baker, 2005). The IEEE 802.15.4 standard was ratified in May 2003 based on Motorola’s proposal. Other companies joined together and formed the ZigBee Alliance in 2002. The ZigBee specification was ratified in December 2004, covering the network, security, and application layers (Baker, 2005).

ZigBee has been designed for low power consumption, low cost, and low data rates for monitoring, control, and sensor applications (Akyildiz, Su, Sankarasubramaniam, & Cayirci, 2002). The lifetime of the networks are expected to be of many months to years with non-rechargeable batteries. The devices operate in unlicensed bands: 2.4 GHz (global), 902-928 MHz (Americas), and 868 MHz (Europe). At 2.4 GHz (16 channels), the raw data rates can achieve up to 250 Kbps, with offset-quadrature phase-shift keying (OQPSK) modulation and direct sequence spread spectrum (DSSS). The 868 MHz (1 channel) and 915 MHz (10 channels) bands also use DSSS, but with binary-phase-shift keying (BPSK) modulation, achieving data rates up to 20 Kbps and 40 Kbps, respectively. The expected range is from 10-100m, depending on environment characteristics.

Each packet, called PHY protocol data unit (PPDU), contains a preamble sequence, a start of frame delimiter, the frame length, and a payload field, the PHY service data unit (PSDU). The 32-bit preamble is designed for acquisition of symbol and chip timing. The payload length can vary from 2 to 127 bytes. A frame check sequence improves the reliability of a packet in difficult conditions. There are four basic frame types: data, acknowledgement (ACK), MAC command, and beacon. The ACK frame confirms to the transmitter that the packet was received without error. The MAC command frame can be used for remote control and nodes configuration.

In 802.15.4 two channel-access mechanisms are implemented, for non-beacon and beacon network. A non-beacon network uses carrier-sense medium access with collision avoidance (CSMA-CA) with positive acknowledgements for successfully received packets. For a beacon-enabled network, a structure called superframe controls the channel access to guarantee dedicated bandwidth and low latency. The network coordinator is responsible for set up of the superframe to transmit beacons at predetermined intervals and to provide 16 equal-width time slots between beacons for contention-free channel access in each time slot (IEEE Std. 802.15.4, 2003; Gutierrez, Callaway, & Barret, 2003).

A ZigBee network can support up to 65,535 nodes, which can be a network coordinator, a full function device (FFD), or a reduced function device (RFD). The network coordinator has general network information and requires the most memory and computing capabilities of the three types. An FFD supports all 802.15.4 functions, and an RFD has limited functionalities to reduce cost and complexity. Two topologies are supported by the standard: star and peer-to-peer, as shown in Figure 2. In the star topology, the
communication is performed between network devices and a single central controller, called the PAN coordinator, responsible for managing all the star functionality. In the peer-to-peer topology, every network device can communicate with any other within its range. This topology also contains a PAN coordinator, which acts as the root of the network. Peer-to-peer topology allows more complex network formations to be implemented, such as the cluster-tree. The cluster-tree network is a special case of a peer-to-peer network in which most devices are FFDs.

**COMPARING ZIGBEE AND BLUE-TOOTH**

Bluetooth and ZigBee have been designed for different applications, and this section makes a comparison between some features of both technologies, such as data rate, power consumption, network latency, complexity, topology, and scalability (Baker, 2005).

In applications where higher data rates are required, Bluetooth always has advantages, especially the 2.0+EDR version (Bluetooth SIG, 2004). While ZigBee mainly supports applications as periodic or intermittent data, achieving rates up to 250 Kbps, Bluetooth can support different traffic types, including not only periodical data, but also multimedia and voice traffic.

ZigBee devices are able to sleep frequently for extended periods to conserve power. This feature works well for energy savings, but increases the network latency because the node will have to awake in order to transmit or receive data. In Bluetooth, the devices do not sleep very often because the nodes are frequently waiting for new nodes or to join other networks. Consequently, data transmission and networks access is fast. Bluetooth devices in sleep mode have to synchronize with the network for communication, while in ZigBee this is not necessary.

The Bluetooth protocol stack is relatively complex when compared to ZigBee. The protocol stack size for ZigBee is about 28 Kbytes and for Bluetooth approximately 100 Kbytes (Geer, 2005). Bluetooth is also more complex if we consider the number of devices. A piconet can have only eight nodes, and a scatternet structure has to be formed to accommodate more nodes (Persson, Manivannan, & Singhal, 2005; Whitaker, Hodge, & Chlamtac, 2005). The Bluetooth SIG does not specify the protocols for scatternet formation. This task is easier in ZigBee networks, since no additional protocols have to be used. In terms of scalability the ZigBee also has some advantages, because network growth is easier to be implemented with flexible topologies. A Bluetooth network growth requires a flexible scatternet formation and routing protocol.

The applications have to consider these characteristics of both protocols when deciding which is the most advantageous for that specific implementation. Bluetooth will fit better in short-range cable replacement, extending LANs to Bluetooth devices and in industries for communication between fixed equipment and mobile devices or machine-to-machine communication (Baker, 2005). ZigBee is most likely to be applied in wireless sensor networks and industries wireless networks, or any other application where battery replacement is difficult and the networks have to live for months to years without human intervention. Many networks may also implement
both protocols in complementary roles using the more suitable characteristic of each for that application. Table 1 shows some features of both technologies.

### RESEARCH CHALLENGES

In the Bluetooth specification there is no information on how a scatternet topology should be formed, maintained, or operated (Persson et al., 2005; Whitaker et al., 2005). Two scatternet topologies that are created from separate approaches can have different characteristics. The complexity of these tasks significantly increases when moving from single piconets to multiple connected piconets.

Some research challenges in Bluetooth scatternets are formation, device status, routing, and intra and inter-piconet scheduling schemes. Each device needs to determine its role with respect to (possibly) multiple piconets, whether master and/or slave. Whitaker et al. (2005) state:

*There is a large degree of freedom in the number of feasible alternative scatternets, which defines a significant combinatorial optimization problem. This is made more difficult by the decentralized nature of the problem, characterized by a lack of a centralized entity with global knowledge.*

The task of packet routing in a scatternet also is not so easy because the packet may have to be transmitted in multiple piconets until it reaches its destination. In a Bluetooth piconet, the master controls the channel access. A slave can send a packet only if it receives a polling packet from the master. Some slaves may participate in multiple piconets, so they become more important than others and the scheduling scheme may give

---

**Table 1. Comparison between Bluetooth and ZigBee**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Bluetooth</th>
<th>ZigBee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data rate</td>
<td>1 Mbps (version 1.2) 3 Mbps (version 2.0)</td>
<td>20-250 Kbps</td>
</tr>
<tr>
<td>Expected battery duration</td>
<td>Days</td>
<td>Years</td>
</tr>
<tr>
<td>Operating frequency</td>
<td>2.4 GHz ISM</td>
<td>868 MHz, 902-928 MHz, 2.4 GHz ISM</td>
</tr>
<tr>
<td>Security</td>
<td>64 bit, 128 bit</td>
<td>128 bit AES</td>
</tr>
<tr>
<td>Network topology</td>
<td>Piconet and scatternet</td>
<td>Star, peer-to-peer, cluster tree</td>
</tr>
<tr>
<td>Protocol stack size</td>
<td>~100 KB</td>
<td>~28 KB</td>
</tr>
<tr>
<td>Transmission range</td>
<td>10-100 meters (depending on power class)</td>
<td>10-100 meters (depending on the environment)</td>
</tr>
<tr>
<td>Network latency (typical)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New device enumeration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Changing from sleep to active</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active device channel access</td>
<td>12 s 3 s 2 ms</td>
<td>30 ms 15 ms 15 ms</td>
</tr>
<tr>
<td>Applications</td>
<td>Cable replacement, wireless USB, handset, headset</td>
<td>Remote control, sensors, battery-operated products</td>
</tr>
</tbody>
</table>
priority for these slaves. The nodes involved in many piconets can only be active one at a time, and the scheduling strategy has to consider this characteristic. As stated in Whitaker et al. (2005), many factors influence the design of scatternet protocols, such as distribution of devices, scalability, device differentiation, environmental dynamism, integration between coordination issues, and level of centralization. The design of efficient protocols could make Bluetooth fit for a wider range of applications.

Although in ZigBee the formation of networks with many nodes is not a great problem, the management of a network with thousands of nodes has not been addressed and may be very difficult (Geer, 2005; Zheng & Lee, 2004). Since ZigBee specification was released after Bluetooth, many important issues have not been addressed, and some distributed protocols will have to be designed for these networks. Both standards have security features, including algorithms for authentication, key exchange, and encryption, but its efficiency still has to be analyzed in networks with many nodes.

Other important issue concerning ZigBee and Bluetooth is the coexistence of both devices, as they use the same 2.4 GHz band, and channel allocation conflicts are inevitable between these WPAN technologies (Chen, Sun, & Gerla, 2006; Howitt & Gutierrez, 2003). This band is also used by wireless LANs based on IEEE 802.11 standard cordless phones and microwave ovens. Interference between near devices may be very common, so coexistence strategies have to be implemented. It is important to study the characteristics of each channel allocation scheme and how each channel allocation scheme interacts with the others. The discussion on the coexistence issue between IEEE 802.11 and the IEEE 802.15-based WPAN technologies has been included in the IEEE 802.15.2 standard.

CONCLUSION

Bluetooth and ZigBee are wireless technologies that may enable many applications of ubiquitous and pervasive computing envisioned by Weiser (1991). Millions of devices are expected to be equipped with one or both technologies in the next few years. This work addressed some of the main features and made some comparisons between them. Some research challenges were described. These issues must be properly studied for the widespread use of ZigBee and Bluetooth technologies.

REFERENCES


IEEE Std. 802.15.1. (2002). *IEEE standard for wireless medium access control (MAC) and physical layer (PHY) specifications for wireless personal area networks*.

IEEE Std. 802.15.4. (2003). *IEEE standard for wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs)*.


**KEY TERMS**

**Carrier-Sense Medium Access with Collision Avoidance (CSMA-CA):** A network contention protocol that listens to a network in order to avoid collisions.

**Direct Sequence Spread Spectrum (DSSS):** A technique that spreads the data into a large coded stream that takes the full bandwidth of the channel.

**Frequency Hopping Spread Spectrum (FHSS):** A method of transmitting signals by rapidly switching a carrier among many frequency channels using a pseudorandom sequence known to both transmitter and receiver.

**Medium Access Control (MAC):** A network layer that determines who is allowed to access the physical media at any one time.

**Modulation:** The process in which information signals are impressed on an radio frequency carrier wave by varying the amplitude, frequency, or phase.

**Pervasive Computing:** An environment where devices are always available and communicate with each other over wireless networks without any interaction required by the user.

**Scatternet:** A group of independent and non-synchronized piconets that share at least one common Bluetooth device.

**Sensor Network:** A network of spatially distributed devices using sensors to monitor conditions at different locations, such as temperature, sound, pressure, and so forth.

**Wireless Personal Area Network (WPAN):** A logical grouping of wireless devices that is typically limited to a small cell radius.

This work was previously published in *Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar*, pp. 272-276, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 3.2
Augmenting Collaboration with Personalization Services

Christina E. Evangelou  
Research Academic Computer Technology Institute, Greece

Manolis Tzagarakis  
Research Academic Computer Technology Institute, Greece

Nikos Karousos  
Research Academic Computer Technology Institute, Greece

George Gkotsis  
Research Academic Computer Technology Institute, Greece

Dora Nousia  
Research Academic Computer Technology Institute, Greece

ABSTRACT
Collaboration is considered as an essential element for effective learning since it enables learners to better develop their points of view and refine their knowledge. Our aim being to facilitate communities of practice members as learners, we argue that collaboration tools should provide personalization features and functionalities in order to fit the specific individual and community learning requirements. More specifically, we propose a framework of services supporting personalization that being embedded in collaboration tools, can act as catalysts for individual and community learning. The proposed set of services has derived after the careful consideration of a generic learner profile, developed to formalize human actors in settings where learning takes place.

INTRODUCTION
As organizations start to acknowledge the significance of communities of practice (CoPs) in helping them meet their business needs and objectives, new efforts to better facilitate the process of learning in these communities are constantly emerging.
Augmenting Collaboration with Personalization Services

(Quan-Haase, 2005). CoPs, also referred to as “knowledge networks,” is a term commonly used to identify institutionalized, informal networks of professionals managing domains of knowledge (Gongla & Rizzuto, 2001). Such communities are formed by groups of people who share an interest in a domain of human endeavour and engage in a process of collective learning (Wenger, 1998). It is this very process of collective learning that creates bonds between them since such communities are formed by groups of people who are willing to share and elaborate further on their knowledge, insights, and experiences (Wenger & Snyder, 2000). Being tied to and performed through practice, learning is considered of premium value by practitioners for improving their real working practices (Steeples & Goodyear, 1999).

Situated learning in particular, that is learning that normally occurs as the function of an activity, context, and culture, is closely related to the social interactions in the community context. Above and beyond learning situated in explicitly defined contexts such as the school classroom, seminars, or even e-learning approaches, modern learning theories strongly support the value of communities and collaborative work as effective settings for learning (Hoadley & Kilner, 2005). More specifically, they emphasize on collaborative learning work that refers to processes, methodologies, and environments, where professionals engage in a common task and where individuals depend on and are accountable to each other. When speaking about collaborative learning, we espouse the Wenger’s perspective of learning as a social phenomenon in the context of our lived experience of participation in the world (Wenger, 1998). As regards to it, an especially valued activity involves information exchanges in which information is constructed through addition, explanation, evaluation, transformation, or summarising (Gray, 2004; Maudet & Moore, 1999). Discoursing, in particular, is considered as an essential element for effective learning, since it enables learners to better develop their points of view and refine their knowledge. This is because, in discoursing, participants focus on the same issues, share their knowledge, and learn to negotiate conflicting opinions in order to reach a commonly accepted solution (Veerman, Andriessen, & Kanselaar, 1998).

Still, it is generally acknowledged that traditional software approaches supporting collaboration are no longer sufficient to support contemporary communication and collaboration needs (Moor & Aakhus, 2006). Research findings on the usage of collaboration tools show that learners are not sufficiently supported in expressing personal ideas and opinions, nor are provided with adequate means for the articulation and sharing of their knowledge. Taking this into account, our work concerns the design of tools that enable discoursing to support collaborative work, emphasis given to aspects, such as the sharing of knowledge and the building of trust. We envisage collaboration tools that can promote learning and encourage creative, parallel, and lateral thinking during collaboration. Towards this, we argue that personalized services can be of great value, as they as they enable the provision of services tailored according to an individual’s (or community’s when applicable) skills, needs, and preferences. In this paper, we present a set of personalization services that has been developed to address the requirements for efficient and effective collaboration between CoP members who can act as catalysts for individual and community learning. Thus, we first performed a comprehensive literature and practice survey of related issues regarding communities of practice, collaboration, and learning. Then, we developed a generic Learner Profile model to formalize CoP members as human actors in settings where learning takes place. The Learner Profile proposed in this paper contributes to the proper user modelling required for the development of virtual environments for collaboration.

The remainder of this paper is structured as follows. The following section discusses
user modelling issues and presents the Learner Profile model of our approach. Next to that, we provide information about the acquisition of the data required for the population of the proposed Learner Profile. Then we present the proposed set of personalized collaboration services towards learning and their relation to the proposed Learner Profile. In the following we discuss implementation issues regarding the embedment of the proposed set of services to collaboration tools. We conclude this paper with final remarks and future work directions.

USER MODELLING

User models are an essential part of every adaptive system. In the following, we discuss design issues, and we present the proposed learner profile model of our approach. The specification of this model is oriented to the development of the personalized services appropriate for learners and/or CoPs.

Design Issues

The primary design aims of our approach in modelling users as learners was to achieve extensibility and adaptability of the user profile as well as the ability to exchange user information between the proposed personalized collaboration services and third-party services. In this context, the proposed learner profile comprises both computational and noncomputational information. Computational information comprises information such as the name, contact details, education, training, and so forth, of users, as well as information about the community they belong to. The noncomputational information is calculated after the processing of the users’ individual behaviour during their participation in system activities. This type of information comprises fields that can be defined during run-time, whenever a new requirement for a new kind of user information is raised. As regards the source of the information stored in the user model, this may derive from the user, the tool, and third-party applications. More specifically, fields that can be filled up by users constitute the user-derived information (e.g., login name, password, address, etc.). In contrast, fields that are calculated and filled up by the tool are machine-derived information (e.g., level of participation, average response time, etc.). Furthermore, some fields can be filled up both from the user and machine (preferences, resources, etc.). In addition, there can be fields that are calculated by external or third-party tools (or applications). Although user and machine-derived information can be easily gathered, third-party tools have to be aware of the user profile and the communication means with the tool in order to interchange data of the user profile. For this purpose, the user profile template is available through an xml schema definition to third-party requestors via Web services. In the storage layer, user records are stored in a relational database and manipulated through SQL queries.

The Learner Profile

In successful learning organizations, individual learning is continuous, knowledge is shared, and the organizational culture must support learning (Gephart, Marsick, Van Buren, & Spiro, 1996; Marsick & Watkins, 1999). Learning entities transform themselves by acquiring new knowledge, skills, or behaviours in their everyday work, through study, instruction, or experience. That is why software tools facilitating working practices for individuals, communities, or organizations should also be conceived and designed as environments where learning takes place. In our approach, collaboration tools should satisfy the community members’ needs to construct and refine their ideas, opinions, and thoughts in meaningful ways, in order to successfully assist individual and community learning. At the same time, individual standpoints should be articulated in such a way that can be proven useful for the rest
of the community’s members. In addition to that, support should be offered for the development of learning skills, such as the interaction with other actors, as well as growth of the learners’ autonomy and self-direction. Moreover, identification of CoP members’ individual characteristics, as well as the culture, norms, and incentive schemes of the community should be appropriately handled.

Research findings about learners’ modelling prove that due to the complexity of human actors and the diversity regarding the learning context, the development of a commonly accepted learner profile is a highly complex task (Dolog & Schäfer, 2005). For instance, the learner model proposed in Chen and Mizoguchi (1999) depicts a learner as a concept hierarchy, but it does not refer to issues such as the learning object, or the learners’ interactions with their environment and other people. However, it provides interesting information about a learner’s cognitive characteristics, and it provides a representation of knowledge assessment issues. Another related approach, the “PAPI Learner” conceptual model, comprises preferences, performance, portfolio, and other types of information (PAPI, 2000). Yet, this model is too generic, as its primary aim is to be portable in order to fit a wide range of applications, and it does not provide any information about a learner’s profile dynamic aspects. The IMS Learner Information Package specification (IMS LIP, 2001) is a useful collection of information that addresses the interoperability of Internet-based Learner Information Systems with other systems that support the Internet learning environment.

After the careful consideration of the above, we developed a generic Learner Profile (see Figure 1) that can be employed for the representation of both individuals and communities as learners (Vidou, Dieng-Kuntz, El Ghali, Evangelou, Giboin, Jacquemart, & Tifous, 2006). The proposed model can be used for developing customized services for both individual and group learners. More specifically, the proposed Learner Profile consists of two types of information, namely static information and dynamic information in compliance with the computational and noncomputational data presented. Static information is considered as domain independent in our approach. The Learner Profile dynamic information elements were chosen to reflect one’s individual behaviour during his participation in a specific CoP’s collaboration activities. Thus, all four dynamic elements, that is, preferences, relations, competences, and experience are to be implicitly or explicitly defined through the learner’s interaction with a tool supporting collaboration. Preferences regarding the use of resources and services provided by the tool, as well as relations among individuals, CoPs,

*Figure 1. The proposed Learner Profile*
and learning items (e.g., argument, URL, or document) can reveal the learners’ different personality types and learning styles. Competences refer to cognitive characteristics such as the creativity, reciprocity, and social skills. Experience reflects learners’ familiarity and know-how regarding a specific domain. It should be noted that all dynamic elements of the proposed Learner Profile can be of assistance towards learning. Nevertheless, the domain of the issue under consideration is a decisive factor. Thus, dynamic aspects of a learner’s profile are treated as domain specific in our approach.

**ACQUIRING THE LEARNER PROFILE DATA**

In order to enable the operation of personalized collaboration services, the Learner Profile has to be populated with the appropriate data. Such data can be acquired in two ways: explicitly from the users’ preferences, and implicitly based on the users’ behaviour while using the system. Static information of the Learner Profile is explicitly provided by the user as a required initialization step of the registration procedure. While such information is usually provided when registering to the system, users should be able to edit this set of profile information at any time. Such explicit data acquisition constitutes a subjective way of profiling, since it depends on the statements made by the user (e.g., experience level, competences, etc.). Their subjective nature may influence personalization services in an unpredictable way (e.g., suggesting to a novice user a document that requires advanced domain knowledge because the user misjudged his experience or competence level). To cope with such issues, we are currently in the process of designing methods that assess explicitly stated profile data, based on the users’ behaviour. We refer to these ways as implicit or behaviour-based data acquisition. In general, the aim of implicit or behaviour-based data acquisition is to assess experience, domains, competences of an individual user based on his behaviour. Implicit data acquisition utilizes the users’ actions and interactions, and attempts to extract information that can permit assessing or augmenting a user profile data.

A special part of the system’s architecture should be dedicated to support implicit data acquisition and interpretation. It consists of a number of modules, each of which is responsible for a particular task (see Figure 2). More specifically, the User Action and Tracking module is responsible for observing user actions and recording them in a special repository of the infrastructure called the Action and Event Store. The Action and Event Store only maintains all actions and events that are useful for implicit user action analysis and does not interpret them in any way. Analysis and interpretation of the gathered data as well as triggering of the appropriate computations (i.e., system reactions) is the main responsibility of the Action Interpretation Engine.

![Figure 2. Data acquisition and interpretation structure](image-url)
Interpretation Engine analyses the available information in the actions and event store and triggers computations that either update accordingly the user profile or execute a particular action. The interpretation engine can be configured using rules that are also stored within the infrastructure, making the interpretation engine rule-based. A rule essentially specifies under which circumstances (i.e., the events and actions of a particular user in the store) an action is triggered. The rule-based nature of the interpretation engine makes the engine itself extensible so that even more cases of implicit data acquisition and interpretation are able to be supported.

Based on the explicit or implicit data, explicit or implicit adaptation mechanisms can be supported within the collaboration tool. Explicit adaptation mechanisms refer to approaches where the tool adapts its services based on the explicitly stated characteristics or preferences of the user. Users are usually aware of explicit adaptations since they themselves triggered the initiation and presence of the respective services. On the other hand, implicit adaptation mechanisms refer to approaches that adapt the system’s services to the user, based on his/her actions within it. Such mechanisms work in the background; Users are usually unaware of the origin of these services since they did not explicitly initiate their activation and, thus, do not perceive their operation. Implicit personalization mechanisms are automatically triggered by the system utilizing implicit or behaviour-based data in the proposed Learner Profile.

In order to enable the foreseen functionalities (such as dynamic update of user information, adaptation of the tool according to the user needs, etc.), the most important actions of the entire set of users’ actions should be tracked down. As regards the User Action Tracking Mechanism, the recorded data about user actions contain information about who did the action, when, what type of action was executed, and what objects were affected by the action. In this way, it will be possible for the system to give valuable feedback to other mechanisms so as to be able to both examine and calculate dynamic user characteristics. Furthermore, a variety of statistical reports that cover both the overall and the specific views of usage of the system should also be produced.

Furthermore, a rule-based approach has been chosen so as to facilitate incorporation of new rules once they are observed, or modification of existing ones if they prove to be too restrictive or even harmful. More specifically, we propose the development of a set of rules that deal with resource access, as access to resources are logged, and a number of rules operate on the logged data to provide additional information to resources and/or user profiles. These can be based on the frequency of access, as well as the competence and experience levels of users (e.g., a document that is frequently accessed by novice users should augment the documents metadata with elements that mirror this fact, so that this document can be recommended to any novice user entering a discussion). A second set of rules observing discussion contribution could control how user behaviour in the context of discussions will affect the users’ competence and experience (e.g., users that actively and frequently participate can be assigned with a high experience level). Another useful indicator associated to the proposed learner profile is the reasoning about how a competence level of a particular user changes in time. This may provide useful insights about the learning capabilities of the particular user and the usefulness of the system.

### PERSONALIZED SERVICES FOR COLLABORATION TOOLS

One of the major challenges in developing software is that all users are different, in that they vary in terms of intelligence, knowledge, training, experience, personality, and cognitive style. Therefore, collaboration tools should provide a set of personalized services, with the emphasis
given to the individuals’ learning styles. In the following we present a set of services employed for enhancing software tools supporting collaboration towards learning. The proposed set of services has resulted out of a thorough investigation of the related literature, existing case studies that consider diverse aspects of learning within communities, as well as a transversal analysis of a set of interviews with real CoP members engaged in various domains of practice.

**Awareness**

According to the findings of our research, CoPs’ members consider system awareness services as the most helpful ones for collaboration tools. Participation awareness provides information about CoP members, online members as well as the discourse moves of individual CoP members. Users will be able to see which user is online, how the space changed by a particular member, and so forth. Social awareness provides information on how members are related to other members in the CoP, and includes statistics about how and how many times members within a CoP communicate with each other and social networks representing the community. Based on the data populated in the Learner Profile, personalized services can provide the proper set of notification actions for the provision of helpful personalized information about system events to CoP members. For instance, a collaboration tool could alert users about the entrance of another user to the system, or about new content insertion into the system.

In order to enable this personalized awareness, terms such as “related” or “interesting” that define a relation between the user and the content should be determined by the user himself, or automatically by the system through the manipulation of some characteristics from the user profile. Furthermore, system awareness can play an important role to assist the familiarization on the new learners of the system. By both informing the CoP moderator about the entrance of a new member and proposing some starting guidelines to the incomer, this service can assist the learning of the way of participation within a CoP. On the other hand, the awareness can provide the moderators with the activity monitoring service that helps the moderator to better understand and manage the whole CoPs’ procedures. That, in turn, contributes to the process of learning the CoP’s moderator role.

Awareness services can also be of use towards the self-evaluation of the participation of a community member, providing him with valuable feedback about his overall contribution to the community and assisting him in collaborative learning as well as in self-reflecting. Using statistic reports populated according to the Learner Profile, such services can measure the level of the member’s contribution to the collaboration procedure. More specifically, these kinds of services can provide with reports about the actual usage of the resources posted by a member, the citations of their resources, or the actual impact of posts to the overall process. In this way, one can be aware of the overall impression that other members have about his participation.

**Allocation of Resources**

Allocation of resources is another service that being personalized in collaboration tools can facilitate learning activities, especially for autonomous learners. As regards to searching for instance, a Learner’s Profile can provide useful information to rank search resources according to a number of factors, such as the learner’s preferences, or even his/her competence and experience level. In this way, the system will be able to adapt to an individual user’s needs. Moreover, the information about the user’s domains of interest will provide additional information with which a search can be better contextualized, thus leading to more relevant results. Furthermore, reasoning mechanisms could be employed for providing the necessary filtering features for
capturing and reusing the knowledge shared in past collaboration activities. In this vein, filtering and recommendation of content services can further support learning. For instance, some of the attached documents of posted positions that contribute to the strengthening of an argument should be suggested for view to the users according to their Learner Profile. Furthermore, a document library could recommend some documents that are related to a specific learner (e.g., experienced learner’s recommendations or popular documents). Thus, members will be able to extend their knowledge through explicit learning of associated content.

Services for classifying other learners according to their domain of expertise can also assist learning in the community. Such services enable the community members to request for suggestion, find and communicate with their coworkers in a knowledgeable way. Furthermore, if coinciding with a community’s norms and wills, such services could also be used for the assignment of weights regarding the weight of a member’s arguments. In addition, services that keep tracking of the members’ activity contribute to the procedure of learning by example, in which a member can learn during watching another one’s practice in collaborative activities.

**Visualization**

It has been widely argued that visualization of collaboration conducted by a group of people working collaboratively towards solving a common problem can facilitate the overall process in many ways, such as in explicating and sharing individual representations of the problem, in maintaining focus on the overall process, as well as in maintaining consistency and in increasing plausibility and accuracy (Evangelou, Karacapilidis, & Tzagarakis, 2006; Kirschner, Buckingham-Shum, & Carr, 2003). Personalized representation of the associated processes, such as the process of discoursing or knowledge sharing, is an essential feature for tools providing effective environments for learning. Furthermore, personalized visualization of context should provide learners with a working environment that fits to their preferred visualization style. System personalization includes alterations in colours, fonts, and text effects, enabling and disabling pieces of information in the working panel, predefinition of system responses in user actions and so forth. In this direction, taxonomies and classification schemes should be employed, wherever possible, as a means for “guiding” users’ cognition. In any case, it should be noted that there is no panacea for the design of user-friendly interfaces; the related practices should be interpreted, refined, and exploited according to the needs of the different types of learners involved in the particular environment. Appropriate navigation and help tools should also be provided for users with diverse expertise. Adaptive User Interfaces should adapt themselves to the learner by reasoning about the user, based on his Learner Profile.

**Building Trust**

Privacy policies and access control services are a critical requirement for the employment of all these services, as well as for the building of trust between the CoP members and the software application. These should be provided in order to satisfy the learner/users’ need to know what information about them is recorded, for what purposes, how long this information will be kept, and if this information is revealed to other people. Furthermore, the security assurance, while establishing connections between users and services, or while accessing stored information, should be taken into consideration as well. Towards this end, two major techniques are broadly used to provide denial of access to data, that is, anonymity and encryption. Anonymity cuts the relation between the particular user and the information about him/her, while information encryption provides protection of the exchanged personal data. In our approach, we
employed the Platform for Privacy Preferences Project (P3P) approach, a W3C recommendation that supports the description of privacy policies in a standardized XML-based form, which can be automatically retrieved and interpreted by the user client (Cranor, Langheinrich, Marchiori, Presler-Marshall, & Reagle, 2002).

IMPLEMENTATION ISSUES

According to current trends in developing Web-based tools, for reasons such as the reusability of components and agility of services, our approach builds on top of a service-oriented environment. In order to exploit advantages enabled by the Service Oriented Architecture (SOA) design paradigm, the proposed set of services should be based on Web service architecture so as to enable the reusability of the implemented modules, as well as the integration or the interoperation with other services (from external systems). An overall design for the enhancement of tools supporting collaboration with personalized functionality towards learning is depicted in Figure 3. In this approach, we sketch a generic architecture design in which a Learner Profile Service is the basis for the storage and the provision of each learner’s characteristics to a set of proposed services that contribute to the system’s personalization. In order to support extensibility, the learning profile service can be dynamically augmented with new learners’ characteristics during run-time. Furthermore, targeting to the openness of the service, the service can provide the learner profile schema in the form of XML Schema Definition (XSD) in the service requestors. Considering the set of proposed services as nonexhaustive, our approach is open for the addition of new personalized services (see Figure 3, block “New Service”) and can use the Simple Object Access Protocol (SOAP) for both internal and external communication.

CONCLUSION

In this paper, we presented a set of services enhancing CoPs interactions and collaborative work based on a generic Learner Profile model. Our approach concerns an alternative form of online learning with different forms of interaction, and a new way of promoting community building. Its purpose is to aid researchers and developers in the development of personalized collaboration.

Figure 3. The proposed services
systems, that is, tools that adapt their structure and services to the individual user’s characteristics and social behaviour. Our main goal being to support individual and community learning, the proposed set of services is based on personalized features and functionalities. We argue that it can further support learning, as well as the achievement of learning objectives, as it can assist CoP members in the development of learning skills such as the interaction with other actors, growth of their autonomy, and self-direction. Nevertheless, in order to be creatively adapted in CoPs’ everyday practices, the proposed services must fit into the specific culture, norms, and incentive schemes of the community. Our future work directions concern the appropriate handling of these issues as well as the full development of the proposed set of personalization services and its evaluation in diverse CoPs.

ACKNOWLEDGMENT

Research carried out in the context of this paper has been partially funded by the EU PALETTE (Pedagogically Sustained Adaptive Learning through the Exploitation of Tacit and Explicit Knowledge) Integrated Project (IST FP6-2004, Contract Number 028038). Also, the authors thank the anonymous referees for their helpful comments and suggestions on the previous versions of this paper.

REFERENCES


Maudet, N., & Moore, D. J. (1999). Dialogue games for computer supported collaborative ar-
Augmenting Collaboration with Personalization Services


Chapter 3.3
Culture and Technology: A Mutual-Shaping Approach

Thomas Herdin
University of Salzburg, Austria

Wolfgang Hofkirchner
University of Salzburg, Austria

Ursula Maier-Rabler
University of Salzburg, Austria

ABSTRACT

The aim of this chapter is to discuss the mutual influence between culture and technology on a broad inter- and transcultural level. Especially, how does information culture shape the meaning of information, communication, and knowledge, and consequently, the design, spread, and usage of ICTs in certain societies? Vice versa, we are interested in the ways in which the spread and usage of ICTs affect the predominating culture. We aim for a model that incorporates cultural as well as technological factors in order to provide a basis for future ICT research that goes beyond both technological determinism and social constructivism. We believe that new technologies indeed can contribute to more justice in the world in terms of access to knowledge and wealth, if sociocultural factors are taken into account more seriously. Current developments in the context of the UN World Summit on the Information Society raise awareness in this direction. At the same time, we are well aware that the loose notion and imprecise definition of the concept of culture allows for the exploitation of the term in empty political and techno-economical policies. Culture degenerates to an indispensable buzzword in the current ICT debate. This chapter is an attempt to introduce the concept of culture into the socioresponsible ICT research on equal terms with technology, economy, and society.
THE RELATIONSHIP BETWEEN CULTURE AND TECHNOLOGY

How can technology be defined? Technology often is considered a means to a particular end, the means being artificially created, not natural, and something that is not directly necessary for the individual or the end user; it serves, rather, to fulfill the need to produce something that is later to be consumed. However, we use the term in a broader sense. We regard technology as being more than just the sum of such artefacts, which are merely the crystallized, concrete manifestations of human behavioral patterns. A method is the how, the way in which a goal is reached and which involves the use of means. A means is a medium in that it mediates between the starting point and the desired result, regardless of what sort of action is involved. Thus, one could speak of social technology (e.g., psychotherapy) as a technology and not merely of technology as something used for (material) production in a society. So, technology also includes the know-how involved in the use and application of the artefacts. In short, technology embraces the ways and means of acting in pursuit of a goal (Hofkirchner, 1999).

How can culture be defined? Using the same analogy for technology, it could be understood to be an equally artefact-based concept, which is not a means to an end but rather an end in itself. That is to say, it is not in itself an essential of life, but rather something that represents a human desire (i.e., what makes humans distinct from other living beings). Here, too, there is a notion that culture is not only the result of a process but also this very process as it moves toward the goal; that is to say, culture is a characteristic of goal-oriented actions (i.e., the striving toward goals as well as the goals themselves) (Hofkirchner, 1999). It is this notion of culture that we refer to in this chapter.

Are there imaginable connections between culture and technology? The two ideal-typical extreme positions are well-known, each making a single direction of determination (Hofkirchner, 1999).

The first position can be referred to as technological determinism, which postulates the total, or at least dominating, influence of technology on culture. Technology is supposed to develop more or less on its own, pushing social development along as it goes. This may be interpreted positively or negatively. An uncritical opinion of Marxist origin saw social advancement as an inevitable result of technical achievements, just as the ideology of the bourgeoisie justified the progress of the technically possible as socially desirable. This view is opposed entirely by fundamentalists who hold technological development responsible for the loss of important values in society. Neither philosophy accepts the possibility of technological development being influenced in any way. Both ignore the fact that there would be no such development if multinational corporations and national governments were to stop investing in research and development; if there were no economic, military, or political interests to divert their resources into these areas; and if there were no values, morals, or norms that underlay these economic, military, or political interests. The fact that on a micro-level there are countless thousands of engineers constantly involved in technology design, and that on a macro-level managers and politicians decide which technological options are realized, supports the second theory—social constructivism—that technology is constructed deliberately to be a part of society. According to this view, the interests of those groups that dominate the genesis of technology finally are embodied in the technology, which in itself cannot be neutral. Here again, both a critical and an approving variant may be distinguished. While the one bemoans the inability of existing technology to pursue ethically justified, socially acceptable, and peaceful and environmentally sound objectives, the other sees the existing economic, democratic, and human rights structures as the best guarantee
of developing optimal technological options. Both versions neglect the inherent dynamism within technological development.

Do the two theories—technological determinism and social constructivism—together give a realistic view of the relationship between technology and culture? This would mean that two equally matched factors—the technical and the cultural—would not be complete without the other. Furthermore, one also might break away from strict determinism and grant each side a measure of independence, thus denying that one side totally dominates the other. But would we then have a workable proposition to discuss, or would we be reduced to the assumption that one factor partly influences the other but is itself partly influenced by its counterpart? This is a superficial answer. Is it not rather the case that the actions we are talking about, whose dependence on mediating factors we want to stress if we are talking about technology and whose immersion in value judgments we wish to highlight when we are discussing culture, not only have an individual character, but rather, through the availability of technological methods and cultural values on the part of society, acquire a deeply societal nature? The use of technology makes every action no longer unique to any individual person. Technology is based on cooperation, be it in the application of special methods, the implementation of these in specific social areas, their invention and development, or in any situation in which the skills and knowledge of other members of society are required. The same holds true for convictions, value judgments, instructions, standards, behavioral patterns, and the like. These are just as much a part of the context of life of the individual, and they promote certain technological methods but discourage others. Technology makes every technologically mediated action into a socially determined one, and its use is a human characteristic. Technological development is part of cultural development; this means that technology is part of culture, and so their relationship to each other is one of part and whole. Culture is the all-embracing factor in this context.

In each part-whole relationship, the parts are the necessary preconditions for the emergence of the whole but are not the sufficient condition for the complete determination of the result. The whole arises from the parts but then exerts control over them in the form of downward causation; the parts are no longer independent of each other as separate entities but are dominated by the whole. The relation of part and whole in regard to technology and culture is, therefore, as follows: technology has the meaning, the purpose, and the task of functioning as means and method for solving social problems. Social interests, cultural values, norms, and morals are thus in the origin and manifestation of technology in its invention, diffusion, and application in the entire process of its development, as its reason for existence. This, however, is insufficient to enslave technology completely. Technology is ambivalent; sometimes it appears to resist our intentions by wholly or partly failing to do what is wanted of it, and other times it not only fulfills our expectations but goes on to do other useful tasks that originally had not been anticipated. Technology represents potential for the realization of social goals. These technologically realizable goals may correspond to pre-existing goals within society; the practical attainment of these by technological means, however, may cause them to change, at least slightly. It is, of course, also possible that the intended goals may differ from those that can be reached with technological support. In this case, new technology may be developed in order to meet the requirements, or the requirements may, as it were, be adapted to fit the reality of what is technically possible. Realizable goals, therefore, do not always exist at the start of the process but may be discovered as options made available by technology. Whether society decides to pursue these goals on the grounds that they are possible is no longer a question of technology but rather of social decision making (Hofkirchner, 1994).
To conclude, we consider the relationship of technology and culture to be dialectic. A relationship is usually called dialectic if, first, the sides of the relation are opposed to each other; second, both sides depend on each other; and third, they form a relation that is asymmetrical (Hofkirchner, 2004). A part-whole relationship is dialectic since part and whole represents opposites, the whole depends on the parts as well as the parts on the whole, and parts and whole build up a hierarchy in which the different levels cannot be replaced by each other.

Considering this notion of the relationship between technology and culture on a broad and general level, the following section attempts to add further thought in the context of information and communication technologies (ICTs) and culture. The specific meaning of information and communication for different societies, which is predominantly the result of a special culture, determines the meaning and, therefore, the spread and usage of ICTs. Vice versa, ICTs have been developed and will be developed in the future in certain (information and communication) cultures, which leads to the functions and practices of use we are facing when we implement ICTs.

**INFORMATION AND COMMUNICATION CULTURES**

When referring to information and communication cultures, we address the basic significance of having access to information and knowledge and the practices of communication and cooperation in a specific society. The most important consideration involves the relationship between those who have access to information that has a profound effect on the distribution of power of control over flows of information within society. It is assumed that within societies with a strong hierarchical structure, the flow and dissemination of public information is restricted to just a few people, while in more liberal societies, there is a far broader basis for direct access to public information. Furthermore, more hierarchically structured societies are less likely to be expected to adapt to the Internet than liberal societies with a flatter hierarchy (Maier-Rabler, 1995, 2002).

The general attitude toward access to information, toward transparency of structures and processes, and toward empowerment and freedom of expression pertaining to a specific society or state is deeply rooted in traditions and practices of social and cultural conditions. The cultural-social framework of a society is formed mainly by the political-social system, by the legislative system, and particularly by the predominating ethic and religious values. As a result of these diverse dimensions, a continuum between the poles of information-friendly vs. information-restrictive cultures emerges (Maier-Rabler & Sutterlütti, 1992; Maier-Rabler, 1995).

Information-friendly societies foster the development of knowledge throughout all groups of a society by providing equal and universal access to all available public information. In information-friendly societies, people have access to public information, freedom of speech is guaranteed to individuals and institutions, and the concept of universal access is understood as the equitable and affordable access to information infrastructure and to information and knowledge essential to collective and individual human development for all citizens. In information-friendly societies, curiosity is encouraged by education systems, and skills for information retrieval are taught rather than just being fed information. Questions count more than answers, and students are encouraged to research instead of memorize given information (Maier-Rabler, 2002).

The political system in information-friendly cultures is likely to be in a form of communicative democracy within a developed system of civil society. Direct democratic participation is a living practice enjoyed by all social groups. The
legal system is likely to be an information-rich, case-based system in which access to information is vital for legal practice.

The economic system in an information-friendly environment strongly depends on access to information and its dissemination to shareholders and customers. Wealth and success are highly valued, and information on turnovers, revenues, and profits are publicly available. Information-friendly societies experience a great expansion in their limitations, especially through the new information and communication technologies. At the same time, it has become clear that without a capable citizenship and without capable institutions, unintended and even unwanted consequences take place. What is more, the current crises of the stock markets have been due to access to information that neither has been audited nor controlled. On a political level, we face a threat to well-established forms of representative democratic systems through populist political trends. New ways of direct democratic participation turn into the opposite, if utilized by people who have not had the chance to acquire the needed skills. However, in information-friendly societies, the chances to implement successful programs to provide equal chances for all members of society to acquire capabilities (in the context of ICT) are higher than in information-restrictive societies.

If we turn to information-restrictive societies, however, we see that they are characterized by a strong hierarchical order throughout society, leading to fewer chances for social, economic, and cultural movement. In such environments, people obtain access to relevant public information when needed, whereby the information is predominantly defined by the authorities or other higher-ranking institutions or persons within the respective hierarchical system. In such societies, people are accustomed to information simply being provided and not having to actively retrieve it. This attitude characterizes the relationship between citizens and authorities, customers and businesses, the public and the media, and students and teacher. The education system in information-restrictive cultures does not encourage curiosity or question-based learning. The “right” answer is the measure of success. What is right and what is wrong again are defined by authorities in the education system. People are not trained to address their environments and to pose questions critically. These answer-oriented societies are an obstacle for the optimal utilization of new information and communication technologies. Digital communication networks such as the Internet work best with a question-oriented approach that leads to a variety of plausible answers in different contexts. Expecting the right and only answer (as people in information-restrictive societies are trained) leads to predictable disappointments and, therefore, less motivation to get involved in new media.

In information-restrictive cultures, the flow of information between authorities and citizens as well as between businesses and customers follows the push principle, whereby authorities and businesses decide which information is being passed on. In such cultures, the Internet is perceived merely as a new and additional (mass) medium to transfer information to a mass audience. Consequently, a huge amount of information and communication capacities of the Internet simply are left unused. As there are not any geographical, national, or cultural borders within digital communication networks, information and applications from information-friendly cultural environments compete with those from information-restrictive cultures on a global stage.

We assume that information-friendly cultures provide a competitive advantage for their members in the global information society.

**THE HUMAN-CENTERED AND CULTURALLY SENSITIVE ICT ADOPTION PROCESS**

This chapter aims toward a better understanding of ICT adoption processes being dependent
from different information and communication cultures. This process, in most societies, is driven by predominantly techno-economic e-policies that are still striving to overcome the Digital Divide and to foster economic growth by means of ICTs on the macro-level of state policy. This approach has been criticized by various authors in recent years (Preston, 2004; van Dijk, 2005; Warschauer, 2002).

Most critics have in common the need to turn away from techno-deterministic viewpoints to human-centered and culturally sensitive approaches. This also can be characterized as a shift from building infrastructures to creating identities, or from bridging the digital divide to closing the knowledge gap. This means putting the individual in the center of the adoption process of technology; therefore, cognitive, cultural, and social factors must be considered in order to achieve a comprehensive understanding.

Following Mansell (2001) and Garnham (1997), we suggest adopting a rights-based capabilities approach in the ICT adoption process to ensure that people have the possibilities to make informed decisions about the specific ways in which they want to make use of ICTs. Acquiring those capabilities first demands awareness processes on an individual cognitive level. Only when people understand the individual and social implications of ICTs will they be able to make informed choices about their specific usage patterns. The stage when people shift from technology-driven skills to culturally embedded understanding is the stage that brings the ICT adoption process from the macro-level of state e-policy to the micro-level of the individual—an indispensable precondition to bring about the skilled user.

This process requires socially or culturally motivated individuals on the one hand and governments who want to offer a set of alternative choices for their citizens to allow them to achieve whatever new-media-lifestyle they want on the other.

As we have already mentioned, the development of these adoption processes depends strongly on the predominating information and communication culture in a given society. In information-friendly environments, people have a greater chance of developing capabilities in the
context of ICT and, therefore, making informed decisions based on the degree of their involvement with new information and communication technologies.

The following model aims to visualize two dimensions of the ICT adoption process: (1) the stages from access to capabilities and (2) the helical transformation of adoption processes (p) as a result of the mutual relation between technology and culture. Every culturally embedded adoption process leaves the new capable user on an advanced stage that itself is the ground for the access step to technology.

**Model of a Human-Centered and Culturally Sensitive ICT Adoption Process**

The adoption process, which also can be considered the major stage for targeted ePolicy measures, starts with the problems of technology-determined access. We need access to technology in order to make experiences and to trigger the following steps. Unfortunately, many processes get stuck in the access stage; “If they build it, they will come” could be the motive for access-only strategies. Most countries favor this access-dominated strategy, which is predominantly in the interest of the technology industry and, therefore, an industry policy measurement.

The critique of the access-only strategy led to a human-oriented enhancement of the same strategy. People need to have adequate skills in order to use the accessed technology. At first glance, this could solve the problem—not only provide people with technology but also train them to use it. Similar to the access stage, the skills stage also is geared predominantly to the interest of the technology industry; in this case, the big international or global software monopolists. Acquiring skills means dealing with a given technology. The creative potential of people in the context of technology is not addressed (National Research Council, 2004).

A further step has to be taken in order to involve the individual in the process of adopting new information and communication technologies. People must know why they should make use of ICTs and not only how to use them. On the cognitive level, the awareness of technology in the specific cultural sphere has to be raised. Here, there is a cultural translation of technology. Only when people understand the diverse patterns of different practices of ICT usage will they be able to make the informed choices as preconditions for gaining capabilities. And only the capable user will provide the basis for economic growth and competitiveness for which most countries, regions, and cultures are striving.

The capable user is the point of departure for the next iteration of the ICT adoption process (p'). Capable users have different demands for access to new technology and also represent a different level for skills training. Such qualified users, who differ in terms of cultural and social backgrounds, represent the input into p'', and so forth.

**DIGITAL CULTURES**

**Cultural Shifts: Transculturality**

In recent decades, the concept of interculturality has been very popular and influential in regard to the fairly young discipline of intercultural communication (Leeds-Hurwitz, 1998). In this context, communication was understood to be an action taking place between countries that were perceived as self-contained units. In this traditional definition, cultures are seen as types of autonomous islands that are virtually completely closed-off, which Beck (1997) called metaphorically the “container theory of society” (p. 49). But modern societies are very diverse entities. They contain and incorporate many elements of different origins, and the boundaries between foreign and indigenous cultures get blurred and finally become untraceable. Tsagarousianou (2004) sug-
suggests that diasporas should not be seen as “given communities, a logical, albeit deterritorialized, extension of an ethnic or national group, but as imagined communities, continuously reconstructed and reinvented” (p. 52). Welsch (1999) developed a new approach of connected cultures, which he called transculturality. This approach emerged due to cultures being interconnected and similar lifestyles merging and being assimilated. Cultures cannot be perceived as homogenous units anymore, because they are complex and diverse in themselves. “Cultures today are extremely interconnected and entangled with each other. Lifestyles no longer end at the borders of national cultures, but go beyond these, are found in the same way in other cultures. The way of life for an economist, an academic or a journalist is no longer German or French, but rather European or global in tone” (Welsch, 1999, 197f.).

This also can be observed in the Internet community. People from different countries use a sort of transcultural ideological language in chat rooms and are united by common interests. Even though they come from very different parts of the world, they have more in common with each other than they have with some members of their respective national communities. The mutuality derived from their similar interests prevails over the mutuality derived from nationality.

Enhancing Welsch’s (1999) concept of transculturality, we consider that this concept needs a more focused perspective on the permeability between global and local cultures, which means that transculturality allows the individual to switch between different identities according to current needs, feelings, interests, and demands. People want to belong to a certain group and want to be identified as a member of such a group; they do not want to constantly act, think, and live on a global level. The identity of the self cannot exist only on a global level, and therefore, “the search for identity, collective or individual, ascribed or constructed, becomes the fundamental source of social meaning. … Yet identity is becoming the main, and sometimes the only, source of meaning in an historical period characterized by widespread destructuring of organizations, delegitimation of institutions, fading away of major social movements, and ephemeral cultural expressions” (Castells, 2001, p. 3).

LINKING CULTURE, KNOWLEDGE, AND ICTS

At this point, we introduce the extended concept of culture, which is intertwined with the concept of knowledge with the aim to discuss the correlation between culture, knowledge, and the role of ICTs. This endeavor eventually should lead to an approach that allows us to connect the complex concept of cultures with its impact on various spheres of our respective lives and, therefore, on our identity. Therefore, the term digital culture will be used to describe the model of mutual influence between culture and technology, which we use as a fundamental framework to develop a new understanding of the use of ICTs. This model aims at an understanding of cultural differences in handling information to guarantee a beneficial development of society.

If the concept of transculturality is introduced into the notion of knowledge, there is a rapid increase of global knowledge. ICTs allow direct communication between vast numbers of people with different cultural backgrounds but do not automatically distribute access to knowledge equally. In fact, many citizens cannot gain access to global knowledge or even local knowledge other than to their own knowledge because of their low economic status (digital divide) and their low educational levels (cultural divide). These divides create groups of haves or have-nots, communication-rich or communication-poor, winners or losers in the globalization process. Concerning identities, these divides determine
the different opportunities of switching identity levels. However, the more people are capable of assuming different identities, both on a local and global level, the more they are capable of gaining advantages in the digital culture.

To avoid getting caught in this gap-trap and to guarantee a sort of mutual benefit, we have to find a way to reconcile some aspects of knowledge production and acquisition, which means that global knowledge has to be incorporated locally (pull factor) in order to allow people to benefit from global knowledge on a local level. Also, local knowledge has to be introduced into the cycle of global knowledge production (push factor) in order to make sure that there is an awareness of the existence of this local identity in a global society. Thus, developments in this global society can be influenced with regard to local positions. We face the challenging task of creating a permeable flow of communication that allows for global and local knowledge bases to interact.

**COMBINING ICTS AND CULTURE TO A NEW APPROACH**

As already mentioned, technology and culture influence each other and are mutually dependent on each other. It is, however, important to discuss the terms *culture* and *technology* and the respective points of view. It is equally important to demonstrate that culture and technology influence each other by using the term *digital culture*.

Drawing upon these basic insights, we will discuss the dialectic of shaping, diffusion, and usage of ICTs in societies and different cultural knowledge bases along the following dimensions: content, distribution, and context.

**DIGITAL CONTENT CULTURE**

This indicator refers to the concept of knowledge production or, in other words, how data are converted into knowledge. According to Willke (2004), one has to distinguish between data and knowledge, even though knowledge management often is mistaken for data preparation and data exchange.

In fact, data are nothing but raw material for knowledge, and in the age of ICT, getting ahold of useful data is not difficult. What is difficult, however, is reducing and filtering huge amounts of potentially useful data and converting them into information first by putting them into a broad context that adds relevance to them; knowledge is gained by putting information into a practical context and modifying or creating a practical due to it in order to make the information practically useful (Willke, 2004).
ICTs, like the Internet, can transport and make available huge amounts of data and information. The content to be distributed is taken from this basic range of knowledge. In a metaphorical sense, the Internet can be linked to a sea of information, all of which is useful in principle. Yet, to get ahold of the invaluable essence of relevant information, we have to siphon off all irrelevant information. The focus is on translating data and information into helpful insights that can be used to improve real-life situations by adding practical relevance to the data.

To guarantee the success of knowledge transfer and the adaptation of new knowledge, a transdisciplinary approach, in addition to an interdisciplinary approach, has to be adapted. This means that different scientific approaches are used, but an effort also is made to involve the local community in the process. In that way, one can ensure that the goals are adapted to the local culture, which increases the likelihood of locals accepting them.

There are three main topics that have to be discussed: knowledge management, learning strategies, and educational approaches.

**DIGITAL DISTRIBUTION CULTURE**

The second dimension illustrates the relationship between the channel that is used for information transport and the dissemination of knowledge—the pull/push strategy.

The first aspect is the communication code: if a message is to be transported, it has to be converted into numbers, text/words, and/or pictures (Willke, 2004). There are limits to the amounts and kinds of information that can be transported in a certain channel. This depends on the type of channel as well as the respective circumstances (e.g., legal, technical, environmental, infrastructural) that, in fact, might influence the usage of the channel.

If we distinguish between explicit and tacit (i.e., structured and unstructured) knowledge, we can see how difficult it is to distribute knowledge. While explicit knowledge (represented in documents, databases, products, and processes) is easy to transfer, tacit knowledge “is more dependent on action, context and personal experience, which makes it difficult to formalize and communicate” (Martin, 2003, p. 44).

The next aspect can be observed in culturally influenced communication rituals. Each and every one of our actions is influenced by culture (i.e., norms, values, beliefs), and by performing these actions repeatedly, we permanently reinforce our cultural understanding. A similar cycle of conditioning can be found in technology. We develop technology by drawing upon our cultural understanding. We then use this technology on a daily basis and, thereby, cause it to impact our identity (reinforcement).

This development can be observed with the personal computer. The term *personal* already indicates that this technology was invented in a very individualistic culture. The more time we spend interacting with computers/technology in this way, the more our patterns of thought and conveying knowledge are assimilated to those used by computers/technology. Our way of thinking becomes more abstract, and knowledge is reduced to mere words and graphics, which lead to future inventions being more abstract as a logical result. The term *digital culture* means that we shape our ICTs and are shaped by them in reverse.

The same applies to the Internet, whose basic technology was developed in the academic-military information culture in California in the late 1950s and early 1960s. This implies a certain practice of converting data into knowledge using the Internet as a practical source for information. In similar information cultures, it is clearer how one can use this kind of data. But people from a different cultural setting who have a different concept of knowledge acquisition might not be
able to make adequate use of the Internet. They might not be familiar with the work processes dominating the information culture within which the Internet was developed. Therefore, it could lead to difficulties to connect to and make use of the Internet. Besides, the way the Internet is used might not cohere with their cultural behavior.

**DIGITAL CONTEXT CULTURE**

There are factors that influence culture and technology on a meta-level. The central questions are: What culture do we live in? What culture do we work in? What culture do we act in?

An important indicator is knowledge as a commodity or as a free public good. First, costs are linked to the question of whether knowledge can be circulated freely or whether it should be treated as property (IPR—intellectual property rights; DRM—digital rights management). Second, costs refer to the investment and maintenance of infrastructure (hardware, software, bandwidth), and finally, we have to calculate the costs for educating people to use ICTs successfully and to develop the ability to convert data into knowledge.

Another important indicator deals with power, which can be explained by using the game theory. It seems that most political and economic decisions are based on the zero-sum game theory, which means that any gain for one player represents an equal loss for the other. We have to face the fact that empowering people also means a loss of power for the powerful to some extent. The powerful create new competitors by empowering other people, societies and/or subcultures. This is not so much about unlimited development as it is about creating a situation of equal opportunities.

Content, distribution, and context are influenced by technology as well as culture. It is not enough to focus only on the digital divide but also on the cultural divide, and by using the concept of digital culture, we can develop a kind of empathy with the goal that we have to create inclusion and development as a central value, if we really want to change from a segregated to an inclusive society.

**THE ONE AND THE MANY**

In this respect, it is worth discussing the relationship of the one and the many. Due to global challenges that endanger the species as a whole and that must be met by a single set of intelligently coordinated actions, the partitions of humankind are on the brink of forming a unit on a planetary scale, and many cultures are on the brink of forming one culture. The awareness of this required delicate relationship between the

### Table 1.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Digital Divide</th>
<th>Cultural Divide</th>
<th>Digital Culture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Data, Information “knowing that …”</td>
<td>Knowledge “knowing how …”</td>
<td>Data, Information, Knowledge “knowing why …”</td>
</tr>
<tr>
<td>Distribution</td>
<td>Channels limited to technical possibilities</td>
<td>Inadequacy between text and channel</td>
<td>Sharing and dissemination of knowledge</td>
</tr>
<tr>
<td>Context</td>
<td>Limited to technical Connectivity</td>
<td>Skills Realization Application</td>
<td>Inclusion Awareness Capabilities</td>
</tr>
</tbody>
</table>
one and the many may serve as a normative idea that guides the measures to be taken to advance world society.

The question is how one of the many relates to another one and how the many relate to the oneness that is made up of the manifold. Is the world society to become the common denominator of the various identities? Or is one of the many the only one? Or are the many merely summands of the individual? Or do the many participate in a one that goes beyond them?

The reductionist way of thinking in intercultural discourse is called universalism. Cultural universalism reduces the variety of different cultural identities to what they have in common. Identities are homogenized by a sort of melting pot that was named McWorld (Barber, 2001). Modernism (i.e., the pursuit of human rights, democracy, and capitalism based on the same mode of metabolism carried out by the same technology everywhere) is universalistic—shimmering between a claim to liberal and pompous imperialistic behavior as it is witnessed by its adversaries. In either case, it gets rid of the richness of cultural identities, the many are reduced to a shallow one, and there is no diversity in the unity.

A second strand in intercultural discourse revolves around the way of thinking that overuses projection. It may be called particularism or totalitarianism. Cultural particularism or totalitarianism extrapolates what separates one cultural identity from the rest and construes an imaginary common. It also leads to homogenization. The melting pot in this case, however, was named Jihad (Barber, 2001), because it is the anti-modern fundamentalism that may be a good example for imposing a certain one out of the many on the rest of them. Here, a culture that is accredited with very specific social relations is raised to the level of the ideal, which is to serve as a model for all other cultures to copy. Thus, a specific form is built up to be the general norm. Inasmuch as it is something particular that is raised in this manner, it concerns particularism.

A third way to conceive intercultural discourse is relativism. Cultural relativism rests on the figure of dissociation. By denying any commonality of different cultural identities, it yields fragmentation. The many fall apart. These concepts of multiculturalism and separatism suit postmodern thoughts. Here, each of the many cultures is seen as something with the right to exist and remain free from external interference. Each special case is made into a norm in its own right. Inasmuch as it is one of many that is made into a norm, we may speak of pluralism. Inasmuch as every special case is treated thus, we must, however, speak of indifferentism. Relativism does not claim general validity and does not wish to unify anything or anyone. The postmodernist form leaves differences as they are. World society would simply be diversity without unity.

None of these three options can satisfy. Either the one is regarded as the necessary and sufficient condition for the many, the many are considered necessary and sufficient for the one, or one and many are deemed independent.

Cultural thinking that reconciles the one and the many is achievable only on the basis of an integration and differentiation way of thinking. It integrates the differences of the manifold cultural identities and differentiates the common as well. Welsch (1999) coined the term transculturalism and notions of glocalization (Robertson, 1992) or new mestizaje (a term coined by John Francis Burke in “Reconciling Cultural Diversity With a Democratic Community: Mestizaje as Opposer to the Usual Suspects” in Wieviorka (2003), which are useful in this context.

The process of emergence of a new sustainable world society may be sketched in terms of dialectics. Diversity is not abolished but rather sublated and leads in an evolutionary leap to a unity through diversity, which, in turn, enables and constrains diversity in order to make it diversity
Culture and Technology

in unity, which then builds the new base for unity through diversity. World culture is located on the macro-level; the partitions of world culture that are located on the micro-level take care of the world culture in order to preserve humanity.

CONCLUSION

Starting with a critique of both techno-deterministic and social-constructive approaches toward the relationship between technology and culture, we argue for a dialectical, mutual-shaping approach. Especially in the context of information and communication technologies (ICTs) and society, this dialectical relationship between culture and technology is important. To strive for the capable user, cultural dimensions have to be incorporated into a model that transfers the spread and usage of technology on the one hand and the social shaping of technology on the other. The concept of digital culture represents a framework that embraces the techno-cultural dimensions of content, distribution, and context. This framework provides an applicable instrument that allows addressing the important questions in the context of technology and society, such as equal knowledge distribution, provision of capabilities, and social inclusion.

REFERENCES


Mansell, R. (2001). New media and the power of networks [Inaugural Professorial Lecture, Dixon’s Chair in New Media and the Internet]. London School of Economics and Political Sci-


Chapter 3.4
Importance of Interface Agent Characteristics from End-User Perspective

Alexander Serenko
Lakehead University, Canada

ABSTRACT

This article reports on an empirical investigation of user perceptions of the importance of several characteristics of interface agents. Interface agents are software entities that are incorporated into various computer applications, including electronic mail systems. As evidenced by the growing body of empirical studies and the increasing number of interface agent-based applications on the software market, there is a strong need for the development of this technology. According to a meta-review of agent-related literature by Dehn and van Mulken (2000), there are several characteristics of interface agents that require special attention from agent developers. However, prior to this study, the importance of these characteristics from the end-user perspective remained unclear. In order to identify the significance of these characteristics, a group of actual users of an e-mail interface agent was surveyed. The results indicate that information accuracy and the degree of the usefulness of an agent are the most salient factors, followed by user comfortability with an agent, the extent of user enjoyment, and visual attractiveness of an agent. The implications of the findings for both theory and practice are discussed.

INTRODUCTION

To create an artificial being has been a dream of men since the birth of science. Professor Hobby (William Hurt) in “Artificial Intelligence” (Spielberg, 2002)

For thousands of years, people have thought of someone doing basic tasks for them. That could be a robot, a cyborg, or a well-trained pet. Not until the beginning of the 21st century did it become possible. Now, with the recent development of telecommunications networks and computer technologies, a new type of software application plays the role of virtual assistants that potentially may alleviate some of the problems associated with the employment of software systems. This class of applications often is referred to as intelligent agents, software agents, avatars, or interface agents. As demonstrated by the growing body of academic literature and by the increasing number of agent-based software applications on the market, there is...
Importance of Interface Agent Characteristics from End-User Perspective

increased interest in the creation of such software entities. In this article, these software systems are labeled as interface agents.

Interface agents emerged from the recent developments in the field of intelligent agents. The idea of software agents was first introduced by John McCarthy (1956, 1958) and later coined by the MIT Lincoln Laboratory computer scientist Oliver Selfridge. In the 1980s, this concept was explored by agent visionaries such as Marvin Minsky and Alan Kay and further utilized in the recent classic works of Pattie Maes, Nicolas Negroponte, Jeffrey Bradshaw, Hyacinth Nwana, and Divine Ndumu. The past few years have witnessed the rapid development of prototypes and working models of intelligent agents, many of which already are incorporated in end-user commercial applications. A number of recent studies demonstrate the fruitfulness and viability of using agent-based technologies in various areas; for example, in automatic negotiation (Castro-Schez et al., 2004; Fatima et al., 2005), natural-language customer support services (Lester et al., 2004), education (Takacs, 2005), and user notification systems (Horvitz et al., 2003). Some academics have shifted their research from human-agent interaction to human-agent cooperation (Rickel & Johnson, 2000; Rickel et al., 2002) and man-machine symbiosis (Klein et al., 2004; Lesh et al., 2004; Lesh et al., 1999), when the human user and the software agent collaborate toward achieving shared goals.

In terms of this article, an interface agent is defined as an autonomous (i.e., independent), continuous (i.e., long-lived), reactive (i.e., it monitors an external environment and reacts to any changes), and collaborative (i.e., it cooperates with other software processes or agents) software entity that exhibits strong visual or audio presence in the computer interface and that communicates with a user directly (i.e., by bypassing intermediaries) (Detlor, 2004; Lieberman & Selker, 2003; Serenko & Detlor, 2004). “Interface agents draw their strength from the naturalness of the living-organism metaphor in terms of both cognitive accessibility and communication style” (Laurel, 1997, p. 68). Typically, interface agents are personalizable and implemented in the form of humanlike or cartoonlike animated characters, electronic figures, graphical user interfaces, textual boxes, or any other visual components (Godoy et al., 2004; Schiaffino & Amandi, 2004).

Having the available agent technology is insufficient; it also should be accepted and utilized appropriately by its target users. For the past 10 years, there have been various attempts to understand what people like or dislike in interface agents and why they adopt or reject them. The goal of this stream of research is to develop a valid, complete list of characteristics that interface agents should possess that would warrant the end-user acceptance of this technology.

By performing a meta-analysis of the human-computer interaction literature, Dehn and van Mulken (2000) presented a comprehensive yet exhaustive list of characteristics of interface agents that potentially may influence the human-interface agent interaction process. Most of these characteristics are drawn from various independent investigations conducted in laboratory settings. At the same time, no study reports how real-life users value the characteristics of an interface agent-based technology. In order to bridge that void, the present investigation attempts to solicit and to analyze the opinions of interface agent users on several key characteristics of the technology. It is assumed that this information potentially may improve the quality of the technology and the way it is delivered to the customer. For example, if agent manufacturers could know what interface agent characteristics are more or less important for users, they would be able to concentrate their short-term efforts to improve positive user perceptions of these characteristics. This, in turn, might increase user satisfaction with agent-based technology and accelerate the rate of innovation diffusion.
As such, Dehn and van Mulken (2000) classified the various characteristics of interface agents (e.g., the user’s subjective experience of the system, the user’s behavior while interacting with the system, and the outcome of the interaction). Each category includes several factors. However, it is not viable to investigate the importance of these characteristics applied to all types of interface agents in a single project. Since interface agents may be incorporated in the form of personal secretaries, Internet guides, electronic commerce assistants, or educators, a separate study is required for each kind of interface agents. It is believed that interface agents embedded in different types of software environments may require certain system-specific features and facets. For example, users who work with an interface agent that facilitates online shopping may look for effectiveness and efficiency. In contrast, people who employ an interface agent as entertainers may emphasize the aspect of enjoyment over that of effectiveness or efficiency.

With respect to the present study, interface agents for electronic mail were chosen for two reasons. First, e-mail is an important telecommunications medium that is utilized heavily by both individuals and organizations. However, today’s e-mail systems provide inadequate support for constantly changing user needs, fail to convey ambiguous content and human emotions, overload people with continually growing flows of unstructured information, and exhibit an inefficient direct manipulation interface. As a result, many individuals feel frustrated utilizing e-mail. The use of interface agents is a potential solution to the currently challenging task of e-mail management. Second, the software market presents several versions of interface agents that have been delivered to end users. Currently, most other types of interface agents have been realized in the form of pilot studies, working prototypes, or beta versions. This identifies the opportunity to reach the actual users of this technology and to poll them directly. It is for these reasons that interface agents for e-mail were selected.

A review of the general characteristics of interface agents presented by Dehn and van Mulken (2000) allowed the identification of several factors that were believed to be applicable to the e-mail environment. Table 1 offers a list of these char-

<table>
<thead>
<tr>
<th>N</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>to believe that an interface agent’s appearance should correspond to its level of intelligence.</td>
</tr>
<tr>
<td>2</td>
<td>to believe that the information provided by an interface agent is accurate.</td>
</tr>
<tr>
<td>3</td>
<td>to like the appearance of an interface agent.</td>
</tr>
<tr>
<td>4</td>
<td>to feel comfortable with an interface agent.</td>
</tr>
<tr>
<td>5</td>
<td>to perceive an interface agent useful.</td>
</tr>
<tr>
<td>6</td>
<td>to perceive an interface agent enjoyable.</td>
</tr>
<tr>
<td>7</td>
<td>to perceive all interactions with an interface agent as natural.</td>
</tr>
<tr>
<td>8</td>
<td>to avoid being distracted by an interface agent while engaged in important tasks.</td>
</tr>
</tbody>
</table>
Importance of Interface Agent Characteristics from End-User Perspective

characteristics. However, little is known about how important these characteristics are for the actual users of e-mail interface agents. As noted by Dehn and van Mulken (2000), the results of the empirical studies that identified these characteristics appear to be mixed and inconsistent.

To bridge this void and to rank the importance of the previous characteristics, this study polled the actual users of e-mail interface agents. It was believed that the end users who have utilized this technology for a long period of time may present valid and reliable information that will be of interest to agent researchers and developers. The following research question was proposed:

*How important are the characteristics of e-mail interface agents identified in Table 1 from the end-user perspective?*

**METHODOLOGY AND RESULTS**

In order to answer the study’s research question, a survey of current and past users of an interface agent-based application for e-mail was conducted. Despite the extensive work underway in the incorporation of interface agents in e-mail applications, most previous studies and projects have been realized in forms of conceptual discussions, preliminary empirical investigations, and pilot systems (Bergman et al., 2002; Dabbish et al., 2005; Florea & Moldovanu, 1996; Griss et al., 2002; Gruen et al., 1999; Lashkari et al., 1994; Maes, 1994; Voss, 2004) rather than in end-user products. E-mail notification applications are one of the first commercial systems that utilize interface agent technologies in the electronic mail environment. This type of interface agents was chosen to conduct a user survey. Out of all commercially available interface agent systems for e-mail, Blind Bat Software was chosen randomly by the researcher, the executives of the company were approached, and agreement was reached. The list of customers who potentially might serve as the study’s participants was sent to the researcher. Figure 1 presents a screenshot of the software product.

In order to poll e-mail agent users on their perceptions of the importance of the characteristics of interface agents, a survey instrument was designed. The questionnaire provided basic instructions, a definition of an interface agent for e-mail, and a list of characteristics. The participants were asked to rate the importance of each characteristic on a scale of 1 to 5, with 1 being not important and 5 being very important.

![Figure 1. E-mail announcer developed by Blind Bat Software (http://www.blindbat.com)](image-url)
and several screenshots of the agent developed by Blind Bat. Users were asked to indicate their opinion on perceptions of the importance of agent characteristics outlined in Table 1. Particularly, the question stated, “Based on your experience with interface agents for e-mail, how important is it for you?” After this, eight statements were provided, measured on a seven-point Likert-type scale ranging from totally unimportant to very important. In addition, demographic information was solicited. The data for this investigation were collected as part of a larger project conducted by Serenko (2005).

By utilizing the total design method (Dillman, 1999), the four-phase survey process was developed. As such, all respondents were e-mailed an initial request to participate in the study and then three follow-up reminders. Fifty-nine usable responses were obtained. An acceptable response rate was achieved. Note that the actual response rate may not be revealed as per the nondisclosure agreement with Blind Bat Software.

Eighty percent of the surveyed users were male, and 20% were female. Over 65% of all users were between 31 and 50 years old, and the 46-to-50-age category was the most frequent user group. Over one-half of the respondents were occupied in the information technology sector; most of them were well-educated and financially well-off and demonstrated a high degree of personal innovativeness in the domain of IT. According to Rogers (2003), this group of people corresponds to innovators, who constitute 2.5% of all people that adopt a particular product or service.

Recall that respondents were asked to rate their perceptions of the importance of eight characteristics of interface agents on a seven-point Likert-type scale. The purpose was to understand what characteristics were more or less imperative from the end user’s point of view. Figure 2 visualizes the results, and Table 2 presents the list of questions sorted by the mean.

To analyze whether there were differences in these eight means, the ANOVA test was conducted. The goal of this statistical method is to determine the existence of differences among several population means (Aczel, 1996). This technique is an extension of the two-sample \( t \) test. The results demonstrated that there was a high degree of confidence that at least some of the means differed from one another (\( F = 12.846, \) d.f. between = 7, d.f., within = 456, significance

Figure 2. User perceptions of the importance of interface agent characteristics

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>appearance vs intelligence</td>
<td>4.22</td>
</tr>
<tr>
<td>natural interactions</td>
<td>5.34</td>
</tr>
<tr>
<td>non-distraction</td>
<td>5.47</td>
</tr>
<tr>
<td>attractiveness</td>
<td>5.78</td>
</tr>
<tr>
<td>enjoyment</td>
<td>5.86</td>
</tr>
<tr>
<td>comfortability</td>
<td>5.90</td>
</tr>
<tr>
<td>usefulness</td>
<td>6.05</td>
</tr>
<tr>
<td>info accuracy</td>
<td>6.28</td>
</tr>
</tbody>
</table>
Importance of Interface Agent Characteristics from End-User Perspective

level = 0.000). To measure the practical value of the detected differences, the effect size was calculated as the ratio of sum of squares between the sum of squares total. The effect size was very strong ($\eta^2 = 0.16$).

After it was determined that differences existed among the means, the Tukey Honestly Significant Difference test was done by using SPSS. The Tukey post hoc test is a statistical method of pairwise comparisons of the population means. It allows the comparison of every possible pair of means using a selected single level of significance. The test yielded a matrix where asterisks (*) indicated significantly different group means at an alpha level of 0.1. Table 3 presents the results of mean comparisons.

Based on these results, several statistically significant differences in item means were observed. Overall, the means of the questions positioned on the left-hand side and right-hand side of Figure 2 strongly differed from one another. This demonstrated a strong degree of confidence that respondents were able to distinguish among the questions and that the results presented in Figure 2 were statistically sound.

**DISCUSSION AND CONCLUSION**

Recall that the purpose of this study was to obtain strong empirical evidence on the importance of interface agent characteristics to bridge the gap in the human-agent interaction literature. The results of the survey showed that trust in an agent (i.e., information accuracy) as well as an agent’s utility (i.e., the persona effect) were the most important factors from the end user’s point of view. They were followed by the degree of conformability and enjoyment with an agent.

First, agent users believed that the accuracy of any information provided by an agent was the most critical factor. This finding is consistent with prior research that points out the importance of trustworthiness in human-agent interaction (Bickmore & Cassell, 2005; Bickmore & Picard, 2005; Hertzum et al., 2002). Indeed, in order to delegate tasks to an agent, a person must believe that the agent will perform them accurately and report back the true rather than the desirable state.

Second, respondents indicated the significance of an agent’s usefulness. This, again, is consistent with prior empirical research and speculations on the importance of the persona effect in agents. The persona effect emerges when an interface agent adds the positive perceptions of usefulness, ease of

<table>
<thead>
<tr>
<th>Based on your experience with interface agents for e-mail, how important is it for you:</th>
<th>Mean</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>to believe that the information provided by an interface agent is accurate?</td>
<td>6.28</td>
<td>1.04</td>
</tr>
<tr>
<td>to perceive an interface agent as useful?</td>
<td>6.05</td>
<td>1.13</td>
</tr>
<tr>
<td>to feel comfortable with an interface agent?</td>
<td>5.90</td>
<td>1.10</td>
</tr>
<tr>
<td>to perceive an interface agent as enjoyable?</td>
<td>5.86</td>
<td>1.13</td>
</tr>
<tr>
<td>to like the appearance of an interface agent?</td>
<td>5.78</td>
<td>1.17</td>
</tr>
<tr>
<td>to avoid being distracted by an interface agent while engaged in important tasks?</td>
<td>5.47</td>
<td>1.74</td>
</tr>
<tr>
<td>to perceive all interactions with an interface agent as natural?</td>
<td>5.34</td>
<td>1.36</td>
</tr>
<tr>
<td>to believe that an interface agent’s appearance should correspond to its level of intelligence?</td>
<td>4.22</td>
<td>1.86</td>
</tr>
</tbody>
</table>
Importance of Interface Agent Characteristics from End-User Perspective

Table 3. The Tukey test

<table>
<thead>
<tr>
<th>(I) CHARACTERISTIC</th>
<th>(J) CHARACTERISTIC</th>
<th>Mean Difference (I-J)</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>appearance corresponds to the level of intelligence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-2.05(*)</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>-1.55(*)</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-1.67(*)</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-1.83(*)</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-1.64(*)</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-1.12(*)</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-1.24(*)</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>information accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2.05(*)</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>.50</td>
<td>.486</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.38</td>
<td>.800</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.22</td>
<td>.986</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.41</td>
<td>.719</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.38(*)</td>
<td>.006</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.81(*)</td>
<td>.028</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>attractiveness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.55(*)</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>.50</td>
<td>.486</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.12</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.28</td>
<td>.956</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.09</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.43</td>
<td>.674</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.31</td>
<td>.920</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>comfortability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.67(*)</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>.38</td>
<td>.800</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.12</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.16</td>
<td>.999</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.03</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.55</td>
<td>.353</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.43</td>
<td>.674</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>usefulness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.83(*)</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>.22</td>
<td>.986</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.28</td>
<td>.956</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.16</td>
<td>.999</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.19</td>
<td>.995</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.71(*)</td>
<td>.092</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.50</td>
<td>.275</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>enjoyment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.64(*)</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>.41</td>
<td>.719</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.09</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.03</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.52</td>
<td>.440</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>natural interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.12(*)</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>.93(*)</td>
<td>.006</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.43</td>
<td>.674</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.55</td>
<td>.353</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.71(*)</td>
<td>.092</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.52</td>
<td>.440</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.12</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>little distraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.24(*)</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>.81(*)</td>
<td>.028</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.12</td>
<td>.920</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.43</td>
<td>.674</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.59</td>
<td>.275</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.40</td>
<td>.761</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.12</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>
use, or enjoyment with an existing system. The key outcome of the persona effect is the improvement of existing software applications by embedding interface agents. By emphasizing the importance of an agent’s usefulness, subjects demonstrated that value-added services were the key factors influencing their adoption decisions.

Third, perceptions of the importance of comfortability and enjoyment with an agent were also high. The extent to which a user feels comfortable employing an agent partially corresponds to the ease of use of the agent.

Fourth, items pertaining to non-distraction and the naturalness of interactions received lower scores. Prior work suggests that a user should perceive all interactions with an agent to be natural, and the agent is not supposed to disrupt current user activities. However, this item received a lower score compared with information accuracy, usefulness, comfortability, and enjoyment.

Finally, in contrast to prior research, respondents stated that the appearance of an agent should not necessarily correspond to its level of intelligence. Two assumptions may explain this contradiction. First, highly innovative individuals might wish to utilize an agent that looks maximally intelligent, regardless of its actual degree of intelligence. Second, if users were not satisfied with the agent’s appearance, they easily might install another one, given that there is a variety of cartoon or humanlike agent characters available on the Web. Thus, end users had control over the interface of an agent that reduced their perception of the importance of the agent’s appearance.

These findings are important for both theory and practice. With respect to theory, the investigation discovered some discrepancies between the view of agent researchers and the opinion of real-life users. With regard to practice, it is suggested that agent designers begin emphasizing the more important characteristics of e-mail interface agents in their products. In the short term, they need to concentrate their efforts on the development of interface agents that provide accurate and reliable information and are perceived to be really useful by the end users. After the issues of information accuracy and usefulness are addressed, agent developers may attempt to improve several other characteristics of interface agents. They may improve the degree of user comfortability with the software, increase user enjoyment, and advance the visual appeal of an agent. In the long term, agent manufacturers may want to decrease the degree of an agent’s intrusiveness and facilitate the naturalness of human-agent interactions. However, it is unlikely that they will need to create an interface agent whose appearance would correspond to its level of intelligence. Instead, they should offer a variety of agent interfaces and leave it up to the end users to decide which one to utilize.

Overall, this investigation is one of the first documented attempts to explore the importance of interface agent characteristics by polling the actual users of this technology. The author hopes that other researchers will continue to explore this field that will lead to the creation of really useful interface agents.

REFERENCES


Importance of Interface Agent Characteristics from End-User Perspective


---

*This work was previously published in International Journal of Intelligent Information Technologies, Vol. 2, Issue 2, edited by V. Sugumaran, pp. 49-60, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
INTRODUCTION

Over the last decade, we have witnessed an explosive growth in the information available on the Web. Today, Web browsers provide easy access to myriad sources of text and multimedia data. Search engines index more than a billion pages and finding the desired information is not an easy task. This profusion of resources has prompted the need for developing automatic mining techniques on Web, thereby giving rise to the term “Web mining” (Pal, Talwar, & Mitra, 2002).

Web mining is the application of data mining techniques on the Web for discovering useful patterns and can be divided into three basic categories: Web content mining, Web structure mining, and Web usage mining. Web content mining includes techniques for assisting users in locating Web documents (i.e., pages) that meet certain criteria, while Web structure mining relates to discovering information based on the Web site structure data (the data depicting the Web site map). Web usage mining focuses on analyzing Web access logs and other sources of information regarding user interactions within the Web site in order to capture, understand and model their behavioral patterns and profiles and thereby improve their experience with the Web site.

As citizens requirements and needs change continuously, traditional information searching, and fulfillment of various tasks result to the loss of valuable time spent in identifying the responsible actor (public authority) and waiting in queues. At the same time, the percentage of users who acquaint with the Internet has been remarkably increased (Internet World Stats, 2005). These two facts motivate many governmental organizations to proceed with the provision of e-services via
their Web sites. The ease and speed with which business transactions can be carried out over the Web has been a key driving force in the rapid growth and popularity of e-government, e-commerce, and e-business applications.

In this framework, the Web is emerging as the appropriate environment for business transactions and user-organization interactions. However, since it is a large collection of semi-structured and structured information sources, Web users often suffer from information overload. Personalization is considered as a popular solution in order to alleviate this problem and to customize the Web environment to users (Eirinaki & Vazirgiannis, 2003). Web personalization can be described, as any action that makes the Web experience of a user personalized to his or her needs and wishes. Principal elements of Web personalization include modeling of Web objects (pages) and subjects (users), categorization of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization.

In the remainder of this article, we present the way an e-government application can deploy Web mining techniques in order to support intelligent and personalized interactions with citizens. Specifically, we describe the tasks that typically comprise this process, illustrate the future trends, and discuss the open issues in the field.

BACKGROUND

The close relation between Web mining and Web personalization has become the stimulus for significant research work in the area (Borges & Levene, 1999; Cooley, 2000; Kosala & Blockeel, 2000; Madria, Bhowmick, Ng, & Lim, 1999). Web mining is a complete process and involves specific primary data mining tasks, namely data collection, data reprocessing, pattern discovery, and knowledge post-processing. Therefore, Web mining can be viewed as consisting of the following four tasks (Etzioni, 1996):

- **Information Retrieval—IR (Resource Discovery):** It deals with automatic retrieval of all relevant documents, while at the same time ensuring that the non-relevant ones are fetched as few as possible. The IR process mainly deals with document representation, indexing, and searching. The process of retrieving the data that is either online or offline from the text sources available on the Web such as electronic newsletters, newsgroups, text contents of HTML documents obtained by removing HTML tags, and also the manual selection of Web resources. Here are also included text resources that originally were not accessible from the Web but are accessible now, such as online texts made for search purposes only, text databases, and so forth.

- **Information Extraction—IE (Selection and Pre-Processing):** Once the documents have been retrieved in the IR process, the challenge is to automatically extract knowledge and other required information without human interaction. IE is the task of identifying specific fragments of a single document that constitute its core semantic content and transforming them into useful information. These transformations could be either a kind of pre-processing such as removing stop words, stemming, etc. or a pre-processing aimed at obtaining the desired representation such as finding phrases in the training corpus, transforming the presentation to relational or first-order logic form, and so forth.

- **Generalization (Pattern Recognition and Machine Learning):** Discover general patterns at individual Web sites or across multiple sites. Machine learning or data mining techniques are used for the generalization. Most of the machine learning systems, deployed on the Web, learn more about the
Web mining refers to the overall process of discovering potentially useful and previously unknown information, knowledge, and patterns from Web data. In this sense, it implicitly covers the standard process of knowledge discovery in databases (KDD) and can be considered as a KDD extension applied to the Web (Markellos, Markellou, Rigou, & Sirmakessis, 2004a). Specifically, Web mining can be categorized into three areas of interest based on which part of the Web is mined:

- **Web Content Mining**: Focuses on the discovery/retrieval of useful information from Web contents/data/documents. Web content data consist of unstructured data (free texts), semi-structured data (HTML documents) and more structured data (data in tables, DB generated HTML pages)
- **Web Structure Mining**: Focuses on the structure of the hyperlinks within the Web as a whole (inter-document) with the purpose of discovering its underlying link structure. Web structure data consist of the Web site structure itself
- **Web Usage Mining**: Mines the secondary data derived from Web surfers’ sessions or behaviors and focuses on techniques that could predict user behavior while the user interacts with the Web (Cooley, 2000). Web usage data can be server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data as the result of interactions

Recently, Web usage mining (Srivastava, Cooley, Deshpande, & Tan, 2000) has been proposed as an underlying approach for Web personalization (Mobasher, Cooley, & Srivastava, 2000). The goal of Web usage mining is to capture and model the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages or items that are frequently accessed by groups of users with common needs or interests. Such patterns can be used to better understand behavioral characteristics of visitors or user segments, improve the organization and structure of the site, and create a personalized experience for visitors by providing dynamic recommendations. In particular, techniques such as clustering, association rule mining, and navigational pattern mining that rely on online pattern discovery from user transactions can be used to improve the scalability of collaborative filtering when dealing with clickstream and e-government data.

**WEB MINING TECHNIQUES IN E-PUBLIC SERVICES**

For the implementation and successful operation of e-government, the proper design, which will be the basis in order to receive a series of strategic, administrative, and operational benefits, is necessary. The application of e-government in the public domain can be gradually performed in fourteen (14) levels (Markellou, Panayiotaki, & Tsakalidis, 2003). This allows the unobstructed flow of information from/to the public sector and
gives the possibility not only to the citizens but also to the enterprises (private sector) to acquire better access in the services that state provides. One of these levels is the upgrade of portal with applications adjusted to every user, where Web mining techniques may be applied to improve access to information through the provided e-services (Markellou et al., 2004a; Markellou, Rigou, & Sirmakessis, 2004b).

Specifically, the deployment of Web mining in the e-government domain relates to the analysis of citizen behavior and the production of adequate adaptations. For example, given a specific citizen, the presentation of required information from an e-government portal can be tailored to meet individual needs and preferences by providing personal recommendations on topics relative to those already visited. This process is typically based on a solid user model, which holds up-to-date information on dynamically changing citizen behavior. This enables on-the-fly portal content assembly, addressing exactly what the citizen needs to know without wasting time on topics the user is already proficient or not interested in. The flowchart of this procedure is shortly illustrated in Figure 1. E-government application constructs users’ profiles integrating various sources of data, pre-processes the data and applies Web mining techniques to provide the users with personalized Web experiences.

In order to personalize an e-government site, the system should be able to distinguish between different users or groups of users. This process is called user profiling and its objective is to create an information base that contains the preferences, characteristics, and activities of the users. In the Web domain, user profiling has been developed significantly, since Internet technologies provide easier means of collecting data about the users of a Web site, which in the case of e-government sites are citizens that must be satisfied by the provided services. A user profile can be either static, when the information it contains is never or rarely altered (e.g., demographic information), or dynamic when the user profile’s data change frequently. Such information is obtained either explicitly (e.g., preferences, background, etc.) using online registration forms and questionnaires, or implicitly, by recording the navigational behavior and other users’ actions from server logs, cookies, and so forth. User profiling is extremely useful to G2C (government-to-consumer) and G2B (government-to-business) applications. For example, a public authority, such as a Ministry of Finance and Economy, can customize its information/services concerning chemical and customs procedures to the relevant actors that import chemical goods. Using this aspect, the actors are facilitated in completing the necessary

![Figure 1. Application of Web mining techniques in e-government domain](image-url)
procedures without getting lost in the “maze” of information provided through a multi-field covering e-government Web site.

Another technique that can be used is the one of clustering. Page clustering identifies groups of pages that seem to be conceptually related according to the users’ perception. User clustering results in groups of users that seem to behave similarly when navigating through a Web site. Such knowledge can be used in e-government in order to perform public services segmentation.

Classification technique can be applied after clustering in order to assign a new user to the defined groups. It uses features with high discriminative ability for defining the various profiles, for example, the profile of an active citizen may include the following values: Sex=male, 34<=Age<=40, Job=worker, Education= basic, MaritalStatus=marital, NumberOfChildren=4, and so forth. This knowledge can be used in applying personalization to e-government services and better supporting the needs of the users providing the right information, to the right people, at the right time.

For discovering relations between different types of available information in an e-government environment association rules can be applied. This technique may identify correlations between pages/users/services or other types of items, not directly connected and reveal previously unknown associations between groups of such items with specific similarities. The form of an association rule can be “65% of citizens that their MaritalStatus=marital search e-government portal for information about LifeEpisodes=having a baby” or “40% of the citizens who accessed help desk asked about online filling-in of tax returns and VAT.” For example, the last rule may indicate that this information is not easily accessible or explanatory enough and requires redesign tasks from the portal technical group.

An extension of the previous technique comprises the sequential pattern discovery that can be used for revealing patterns of co-occurrence, which incorporates the notion of time. For example a pattern may be a Web page or a set of pages accessed immediately after another set of pages: “55% of new businesses who apply for a certain certificate will use the certificate within 15 days” or “Given the transactions of a citizens who has not apply for any information/services during the last 3 months, find all citizens with a similar behavior.”

Finally, as search engines often appear as a helpful tool at e-government, personalized Web search systems may be used to enhance their functionality. In order to incorporate user preferences into search engines, three major approaches are proposed (Shahabi & Chen, 2003):

- **Personalized Page Importance:** Modern search engines employ the importance scores of pages for ranking the search results, as well as traditional text matching techniques.
- **Query Refinement:** A process composed of three steps: obtaining user profiles from user, query modification and refinement.
- **Personalized Metasearch Systems:** Metasearch systems could improve the retrieval rate by merging various ranked results from multiple search engines into one final ranked list.

**FUTURE TRENDS**

On the road to enhance an e-government application and treat each user individually, personalization plays a central role. The benefits for both public authorities and citizens are significant when it really works. However, several issues still remain unclear. First of all, determining and delivering personalization is a data intensive task and requires numerous processing steps. This usually causes intolerably long response times, which in turn may lead to site abandonment. To avoid this constrain, parts of the process can be executed
offline or special algorithms and structures can be used to guarantee fast online operation.

Another challenge is to ensure personalization accuracy. It is true that unsuccessful recommendations can slow down the process, confuse, and disorientate users. It is preferable not to deliver any recommendations than deliver a set of useless or harmful ones. Apart from that, personalization should be delivered in the appropriate way (avoiding user intrusion and loss of concentration) and not deprive users control over the whole process. Moreover, as e-government sites are dynamic environments, issues concerning the content or structure updating e.g. newly added topics, pages, services, etc. can be taken into consideration.

Last but not least, privacy violation during the user profiling process should be encountered (Volokh, 2000). Many users are reluctant to giving away personal information either implicitly as mentioned before, or explicitly, being hesitant to visit Web sites that use cookies (if they are aware of their existence) or avoiding to disclose personal data in registration forms. In both cases, the user loses anonymity and is aware that all of his actions will be recorded and used, often without his consent. Additionally, even if a user has agreed to supply personal information to a site, through cookie technology such information can be exchanged between sites, resulting to its disclosure without his permission. Although the new technologies and products for protecting user’s privacy on computers and networks are becoming increasingly popular, none can guarantee absolutely secure communications. Electronic privacy issues in the foreseeable future will become highly crucial and intense (Markellos et al., 2004b).

CONCLUSION

Governments enhance their attempt to offer efficient, advanced and modern services to their users (citizens and businesses) based on information and communication technologies and especially the Web. The remarkable acceptance of this powerful tool has changed the way of conducting various activities and offers citizens, businesses and public authorities limitless options and opportunities. However, the emerging problem to deal with is the way an e-government can provide its users with the right information and service according to their specific needs and preferences. To this direction, Web mining and personalization are used for supporting tailored Web experiences.

These techniques appear as the most promising for the future, since they help to establish one-to-one relationships between users and governments, improve the performance of provided information and services, increase users’ satisfaction and promote e-loyalty. On the other hand, governments take advantage of them, as long as they save costs (e.g., transactions, communication, task management, etc.), improve response times, automate various processes, provide alternative channels of cooperation and communication, and upgrade and modern their profile and image.

Many research and commercial approaches, initiatives and tools are available, based on Web site structure and contents, user’s navigation, behavior and transaction history, server log files, and so forth. However, personalization requires rich data in order to provide successful output. This is not always feasible, since many users are often negative towards the idea of being stereotyped. Moreover, individuals’ privacy has been put in jeopardy by the tasks of recording their activities and saving the appropriate data into their profiles. Summarizing, governments should work hard in the direction of providing the legal framework for ensuring the protection of users’ privacy and also eliminating the possibility of misuse their personal information.
REFERENCES


Madria, S. K., Bhowmick, S. S., Ng, W. K., & Lim, E. P. (1999, August 30-September 1). Research issues in Web data mining. Proceedings of Data Warehousing and Knowledge Discovery, Florence, Italy (pp. 303-312).


Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from Web data. SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, 1(2), 12-23.

Web Mining for Public E-Services Personalization

KEY TERMS

Clickstream: It is a record of a user’s activity on the Internet, including every Web site and every page of every Web site that the user visits, how long the user was on a page or site, in what order the pages were visited, any newsgroups that the user participates in and even the e-mail addresses of mail that the user sends and receives. Both ISPs and individual Web sites are capable of tracking a user’s clickstream.

Cookie: The data sent by a Web server to a Web client, stored locally by the client and sent back to the server on subsequent requests. In other words, a cookie is simply an HTTP header that consists of a text-only string, which is inserted into the memory of a browser. It is used to uniquely identify a user during Web interactions within a site and contains data parameters that allow the remote HTML server to keep a record of the user identity, and what actions she/he takes at the remote Web site.

Data Mining: The application of specific algorithms for extracting patterns (models) from data.

Knowledge Discovery in Databases (KDD): The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Server Log: Web servers maintain log files listing every request made to the server. With log file analysis tools, it’s possible to get a good idea of where visitors are coming from, how often they return, and how they navigate through a site. Using cookies enables Webmasters to log even more detailed information about how individual users are accessing a site.

Web Mining: The discovery and analysis of useful information from the Web.

Web Personalization: It is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user’s navigational behavior (usage data) in correlation with other information collected in the Web context, namely structure, content and user profile data. Due to the explosive growth of the Web, the domain of Web personalization has gained great momentum both in the research and the commercial area.

Web Usage Mining: The application of data mining techniques to Web clickstream data in order to extract usage patterns.

This work was previously published in Encyclopedia of Digital Government, edited by A. Anttiroiko and M. Malkia, pp. 1629-1634, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
INTRODUCTION

The decreasing cost of networking technology and network-enabled devices is driving the large scale deployment of such networks and devices so as to offer many new and innovative services to users in ubiquitous computing. For example, when you carry your mobile laptop or personal digital assistant (PDA) around, or drive on the road, various services have been made available, ranging from finding a local printer to print a file, to instantaneously knowing about the traffic situation from traffic-cameras and other sensors along a highway.

To achieve the above, every participating network-enabled end-device must solve an interesting technical problem, i.e., to locate a particular network service or device out of hundreds of thousands of accessible services and devices. Such service advertising and discovery is important as mobile devices and mobile wireless devices proliferate on networks. For this reason, a service discovery and advertising protocol is an
important tool to help these devices find services on the network wherever they connect, and to let other network users know about the services they are offering.

Context-aware service discovery, on the other hand, would help users to find services that are most appropriate based on fast-changing client conditions, such as location. For example, most laptops are statically configured to print to dedicated office printers. With the help of the context-awareness, a laptop could find the nearest accessible printer attached to the network that the laptop is currently plugged into.

**SERVICE DISCOVERY PROTOCOLS**

As new services are made available, clients would be able to obtain services based on a set of client-defined keywords or attributes. These attributes would allow the client to narrow the scope of the search so that only those services that pertain to its needs would be discovered. Furthermore, the client would be able to automatically start interacting with the newly discovered services without any programming (Intel).

For the past few years, competing industries and standards communities have been hotly pursuing the technologies for service discovery. E-speak (HP, 2002), UDDI (Universal Description, Discovery, and Integration), Sun’s Jini (Arnold, ed. 2001), Microsoft’s Universal Plug and Play (UPnP), IBM’s Salutation (Salutation Architecture Specification), Service Location Protocol (SLP) of IETF (Guttman, 1999), and Bluetooth’s Service Discovery Protocol (SDP) (Helal, 2002) are among the front-runners in this new race. E-speak and UDDI are designed specifically for discovering Web services. On the other hand, Jini, UPnP, and Salutation are geared toward services furnished by hardware devices such as printers, faxes, etc. We elaborate each of these in detail next.

- **E-speak** is an open software platform designed by HP (Hewlett-Packard Co.) to facilitate the delivery of e-services (electronic services) over the Internet. Based on Extensible Markup Language (XML), the E-speak Service Framework Specification makes it possible for e-services to advertise, discover, negotiate, and form contracts, learn each other’s interfaces and protocols, and invoke each other’s services, all without human intervention.

- **UDDI** protocol is one of the major building blocks required for successful Web services. UDDI creates a standard interoperable platform that enables companies and applications to quickly, easily, and dynamically find and use Web services over the Internet. UDDI is a cross-industry effort driven by major platform and software providers, as well as marketplace operators and e-business leaders within the OASIS (Organization for the Advancement of Structured Information Standards) consortium.

- **Jini** is a distributed service discovery and advertisement architecture that relies on mobile code and leverages the platform independent of Java language. Jini entities consist of services, lookup servers that catalog available services, and clients that require services. All service advertisements and requests go through the lookup servers. To register service availability or to discover services, a service or client must first locate one or more lookup servers by using a multicast request protocol. This request protocol terminates with the invocation of a unicast discovery protocol, which clients and services use to communicate with a specific lookup server. A lookup server can use the multicast announcement protocol to announce its presence on the network. When a lookup server makes an announcement, clients and services that have registered interest in receiving announcements of new lookup services are notified.
• **UPnP** is considered at some levels to be a natural extension of Microsoft Plug and Play to the networking scenario. The UPnP specification describes device addressing, service advertisement and discovery, device control, event, and presentation. AutoIP (Troll, 2000) protocol is used to allow devices to dynamically claim IP addresses in the absence of a DHCP server. UPnP uses simple service discovery protocol (SSDP) for service discovery. SSDP uses HTTP over multicast (HTTPMU) and unicast UDP (HTTPU) to announce a device’s presence to others and discover other devices and services. Each service can be searched by any of its three associated IDs—service type, service name, and location. An UPnP service description includes a list of actions to which the service responds and a list of variables that model the service’s state at runtime. After discovery, a control point sends a message to the service’s specified control object URL (Uniform Resource Locator) according to the Simple Object Access Protocol (SOAP). The device or service returns action-specific values. A control point can also subscribe to receive event notifications which are updates of services’ state in a format of General Event Notification Architecture (GENA).

• **Salutation** is an architecture for service advertisement, discovery, and invocation among devices and services of dissimilar capabilities. The architecture is composed of three fundamental components: functional units, Salutation managers, and transport managers. In Salutation, the concept of a “service” is broken down into a collection of functional units with each unit representing some essential feature. A service description is then a collection of functional unit descriptions, each having a collection of attribute records (name, value, etc.). These records can be queried and matched against during the service discovery process. The Salutation manager contains a registry to keep information about services, and a client can register or un-register itself. An ensemble of Salutation managers can coordinate with one another to exchange registration information, even with those on different transport media. Communication between Salutation managers is based on remote procedure call (RPC). A discovery request is sent to the local Salutation manager, which in turn will be directed to other Salutation managers. This cooperation among Salutation managers forms a conceptually similar lookup service to Jini. One difference, though, is that it is distributed over the network. The transport manager isolates the implementation of the Salutation managers from particular transport-layer protocols and hence gives Salutation network transport independence.

• **SLP** is an Internet Engineering Task Force (IETF) standard for decentralized, lightweight, and extensible service discovery. It uses service URLs, which define the service type and address for a particular service. For example, “service:printer:lp://hostname” is the service URL for a line printer service available at hostname. Based on the service URL, users (or applications) can browse available services in their domain and select and use the one they want. SLP establishes a framework for resource discovery that includes three “agents” that operate on behalf of the network-based software: User Agents (UA), Service Agents (SA), and Directory Agents (DA). UAs perform service discovery on behalf of client software. SAs broadcast the advertisement of the location and attributes on behalf of services. As a centralized service information repository, the DA caches advertisements from SAs and processes discovery queries from UAs.

• **Bluetooth SDP** is specific only to Bluetooth devices (Specification of the Bluetooth System). It defines a very simple service
discovery mechanism for peer-to-peer type networks. It does not provide access to services, brokering of services, service advertisements, or service registration, and there is no event notification when services become unavailable (Helal, 2002). Since there is no centralized directory for service registration, the attributes about a service are stored in the local service discovery server in each mobile device. Clients can search for available services either by service classes, which correspond to unique 128-bit identifiers and uniquely identify types of devices, or by matching attributes. The server replies whether a service is available on the queried platform only and a higher level protocol must provide specific actions to utilize the service.

All of the above mentioned discovery mechanisms have their pros and cons and are valuable within their own infrastructures. The comparison of service discovery protocols is shown in Table 1. Most of them promise similar functionalities which are targeted at reducing configuration hassles, improving device cooperation, and automating discovery of required services. However, due to the lack of a widely adopted service discovery mechanisms, these different products have different APIs and do not work together. In addition, current service discovery frameworks do not consider important context information. Therefore, none of them offer a universal solution to context-aware service discovery.

**CONTEXT-AWARE SERVICE DISCOVERY**

The basic areas we need to address in order to create a context-aware service discovery architecture are:

---

**Table 1. Comparison of service discovery protocols**

<table>
<thead>
<tr>
<th>Developer</th>
<th>E-Speak</th>
<th>UDDI</th>
<th>Jini</th>
<th>UPnP</th>
<th>Salutation</th>
<th>SLP</th>
<th>Bluetooth SDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service type</td>
<td>Web services</td>
<td>Web services</td>
<td>Hardware devices with JVM (Java Virtual Machine)</td>
<td>Hardware devices</td>
<td>Hardware devices</td>
<td>Hardware devices</td>
<td>Only Bluetooth devices</td>
</tr>
<tr>
<td>Central cache repository</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes (lookup server)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Operation w/o directory</td>
<td>-</td>
<td>-</td>
<td>Lookup table required</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Network transport</td>
<td>TCP/IP</td>
<td>TCP/IP</td>
<td>Independent</td>
<td>TCP/IP</td>
<td>Independent</td>
<td>TCP/IP</td>
<td>Bluetooth communication</td>
</tr>
<tr>
<td>Programming language</td>
<td>XML</td>
<td>Independent</td>
<td>Java</td>
<td>Independent</td>
<td>Independent</td>
<td>Independent</td>
<td>Independent</td>
</tr>
<tr>
<td>OS and platform</td>
<td>Independent</td>
<td>Dependent</td>
<td>JVM</td>
<td>Dependent</td>
<td>Dependent</td>
<td>Dependent</td>
<td>Dependent</td>
</tr>
</tbody>
</table>
1. **Interoperability:** Our proposed architecture can interoperate with and provide value added functionality to existing service discovery mechanisms.

2. **Context-awareness:** It will acquire and search for context, and carry out context reasoning to find the most appropriate services for users/applications.

3. **Security, privacy, and trust:** All of this must be done in a manner that provides a sufficient level of security, privacy, and trust to meet the customer’s needs.

Based on the above requirements, we propose a context-aware service discovery architecture as shown in Figure 1. In order to interoperate with existing service discovery mechanisms, we adopt OSGi (Open Service Gateway Initiative) architecture (Marples, 2001; Dobrev, 2002; Kim, 2001) to accommodate the different service collaboration.

Founded in March 1999, the OSGi creates a set of open specifications for the delivery of a wide array of services to end users. The OSGi mission is to enable the deployment of services over wide area networks to local networks and devices (Marple, 2001). The key benefits of OSGi are the following:

- Platform independent. Based on Java technologies, the OSGi gateway can run on most operating systems.
- Multiple network technology support. Home network technologies, such as Jini and UPnP, facilitate device and service discovery in the OSGi gateway (Dobrev, 2002).

![Figure 1. Context-aware service discovery architecture](image-url)
Multiple services support and service collaboration. Multiple applications from different service providers are able to run on a single service platform and provide functionality to other services.

**Interoperability**

The benefits of OSGi provide a feasible interoperability platform for different service discovery mechanisms. Each of the existing service discovery mechanisms has a wrapper to communicate with the OSGi. Wrappers transform the functions of those entities into the form of services and publish those services in the OSGi framework service registry. In the OSGi framework, entities discovered by existing service discovery protocol are represented in the form of services in a standardized way. As the functions of every entity are described using common convention, entities can thus understand each other and collaborate to achieve a certain goal.

**Context-Awareness**

Context information plays an important role in making the physical spaces “smart.” Users and applications often need to be aware of their surrounding context and adapt their behaviors to context changes. Context-awareness involves the use of context to provide relevant services to the user, where relevancy depends on the user’s task (Dey, 2000).

An appropriate model should address different characteristics of context information, such as dependency and uncertainty. Context acquisition is closely coupled with sensors to acquire context data from physical or virtual sensors. In our earlier work, we have proposed an ontology-based context model to describe context information in a semantic way, which exhibits features such as expressiveness, extensibility, ease of sharing and reuse, and logic reasoning support (Zhang, 2006). Context search provides users and applications both synchronous context query service and asynchronous context event notification service. A context processing and management engine called Semantic Space has also been implemented to support context reasoning (Wang, 2004). Context reasoning uses context to carry out logical reasoning to provide the most appropriate services to the users/applications.

**Security, Privacy, and Trust**

In the service discovery, the interaction and information exchange between clients and services must be secure, private, and trustworthy. OSGi defines SSL (Secure Sockets Layer) service for secure network communication and enforces policy-based fine-grained control mechanisms to permit local resource access. The relevant techniques will be exploited and the research works in this area are still ongoing as shown in orange color in Figure 1. We will elaborate more in the future trends.

**FUTURE TRENDS**

The pervasive availability of embedded devices in the environment imposes significant technical and research challenges in service cooperation and interaction. One aspect that captures more attention at the present is security, privacy, and trust which are explained briefly next.

**Security** includes the three main properties of confidentiality, integrity, and availability (Stajano, 2002). **Confidentiality** is concerned with protecting the information/service from unauthorized access; **Integrity** is concerned with protecting the information/service from unauthorized changes; and **Availability** is concerned with ensuring that the information/service remains accessible.

**Privacy** is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information is communicated to others. Privacy is about protecting
users’ personal information. Privacy control relates not only to the process of setting rules and enforcing them, but also to the way privacy is managed/controlled adaptively according to changes in the degree of disclosure of personal information or user mobility from one space to another (Cook, 2005).

**Trust** management is to develop a trust specification that can be analyzed and evaluated before appropriate interactions/transactions really start. As clients and services often interact and collaborate with each other in an ad-hoc manner, those clients and services may come from unfamiliar administrative domains and therefore be completely unknown to each other. To safely take advantage of all the possibilities, it is essential to assess the confidence level of involved parties, estimate the likely behavior of entities, and recommend a certain level for interactions, which is so called trust management. Trust management involves the trust model, the trust and recommendation specifications for different entities from different certified authorities, and risk thresholds. The trust management service uses the trust model to compute the trust value for each interaction. The trust value indicates the degree of confidence and is used as a parameter to determine the access level of services or degree of information interaction or exchange.

From the above elaboration, we may predict that the future trends for context-aware service discovery will involve addressing the issues related to security, privacy, and trust. The techniques related to authentication, authorization, cryptography, and trust management will be exploited for the interaction and information exchange between clients and services.

**CONCLUSION**

This paper surveys the existing service discovery protocols, such as E-speak, UDDI, Jini, UPnP, Salutation, SLP, and SDP, and discusses the different mechanisms. All of the above mentioned mechanisms have their pros and cons, but none of them offer a universal solution to context-aware service discovery. Next, we identified the requirements of service discovery mechanisms and described a context-aware service discovery architecture to support the interoperability and context-awareness which are required in ubiquitous computing. Finally, the future trends for context-aware service discovery will involve addressing the issues related to security, privacy, and trust.

**REFERENCES**


UDDI (Universal Description, Discovery and Integration). http://www.uddi.org/


**KEY TERMS**

**Client:** An application that is interested in or requires some other application to perform some type of work for the client (Intel).

**Context:** Any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and the application themselves (Dey, 2001).

**Context-Awareness:** To use context to provide relevant services to the user, where relevancy depends on the user’s task (Dey, 2000). A **service** is a component or application that performs the work on behalf of a requesting application or client (Intel).

**Service Advertisement:** Is responsible for advertising a given service description on a directory service or directly to other hosts in the network. The effectiveness of an advertisement is measured as a combination of the extent of its outreach and the specificity of information it provides up front about a service (Sen, 2005).

**Service Description:** Is responsible for describing a service and the type of context information in a comprehensive, unambiguous manner that is machine interpretable to facilitate automation and human readable to facilitate rapid formulation by users (Sen, 2005).
Context-Aware Service Discovery in Ubiquitous Computing

Service Discovery: Is the keystone and carries out three main functions. It formulates a request, which is a description of the needs of a service consumer. This request is formatted in a similar manner to the service description. It also provides a matching function that pairs requests to services with similar descriptions, so that the service which can best fulfill service consumer’s needs on peer devices is selected. Finally, it provides a mechanism for the service consumer to communicate with the service provider (Sen, 2005).

Service Invocation: Is responsible for facilitating the use of a service. Its functions include transmitting commands from the service consumer to the service provider and receiving results. It is also responsible for maintaining the connection between the consumer and the provider for the duration of their interaction. A good invocation mechanism abstracts communication details from the service consumer and, in the case of network failure, redirects requests to another provider or gracefully terminates (Sen, 2005).

Ubiquitous Computing (ubicomp, or sometimes ubiqcomp): Integrates computation into the environment, and enables people to move around and interact with information and computing naturally and casually. One of the goals of ubiquitous computing is to enable devices to sense changes in their environment and to automatically adapt and act based on these changes according to user needs and preferences.

This work was previously published in Encyclopedia of Internet Technologies and Applications, edited by M. Freire and M. Pereira, pp. 119-125, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 3.7
Ontological Engineering in
Pervasive Computing Environments

Athanasios Tsounis
University of Athens, Greece

Christos Anagnostopoulos
University of Athens, Greece

Stathes Hadjiiefthymiades
University of Athens, Greece

Izambo Karali
University of Athens, Greece

ABSTRACT

Pervasive computing is a broad and compelling research topic in computer science that focuses on the applications of technology to assist users in everyday life situations. It seeks to provide proactive and self-tuning environments and devices to seamlessly augment a person’s knowledge and decision making ability, while requiring as little direct user interaction as possible. Its vision is the creation of an environment saturated with seamlessly integrated devices with computing and communication capabilities. The realisation of this vision requires that a very large number of devices and software components interoperate seamlessly. As these devices and the associated software will pervade everyday life, an increasing number of software and hardware providers will deploy functionality in pervasive computing environments (PCE). That poses a very important interoperability issue, as it cannot be assumed that the various hardware and software components share common communication and data schemes. We argue that the use of Semantic Web technologies, namely the ontologies, present a intriguing way of resolving such issues and, therefore, their
INTRODUCTION

The vision of pervasive computing presents many technical issues, such as scaling-up of connectivity requirements, heterogeneity of processors and access networks, and poor application portability over embedded processors. These issues are currently being addressed by the research community; however, the most serious challenges are not technological but structural, as embedded processors and sensors in everyday products imply an explosion in the number and type of organisations that need to be involved in achieving seamless interoperability (O'Sullivan, 2003). In a typical pervasive computing environment (PCE) there will be numerous devices with computing capabilities that need to interoperate (Nakajima, 2003). These devices might be of different vendors and may operate based on different protocols. Therefore, the key issue in deploying a PCE is achieving application level interoperability. The complexity of such a venture is considerable. It is extremely difficult to reach agreements when the players involved expand from all the hardware and software providers (e.g., IBM, HP, Microsoft) to all the organisations that will equip their products with computing and communication capabilities (e.g., coffee machines, refrigerators). Therefore, we cannot rely on shared a priori knowledge based on commonly accepted standards to resolve the issue. Instead, software components must adapt to their environment at runtime to integrate their functionality with other software components seamlessly. An intriguing way of resolving this issue is the use of semantics, namely the use of Semantic Web technologies such as ontologies. In this manner, software entities provide semantically enriched specifications of the services that they provide and the way they should be invoked. Moreover, the data that are exchanged are also semantically enriched, enabling the entities to reason and make effective decisions. This is particularly important for the description of contextual information, which is of main interest in a PCE. As context we identify any information that is, directly or indirectly, associated with any entity in the environment.

The novelty of the Semantic Web is that the data are required to be not only machine readable but also machine understandable, as opposed to today’s Web which was mainly designed for human interpretation and use. According to Tim Berners-Lee, the Director of World Wide Web Consortium, “the Semantic Web’s goal is to be a unifying system which will (like the Web for human communication) be as un-restraining as possible so that the complexity of reality can be described” (Berners-Lee, 2001). With the realisation of a Semantic Web it would be easy to deploy a wide range of services that would be almost impossible to manage in the current Web. Semantics enable developers to create powerful tools for complex service creation, description, discovery, and composition. The application areas of the Semantic Web extend from knowledge repositories to e-commerce and from user profiling to PCE.

New standards are being developed as a first step in realising the Semantic Web. The Resource Description Framework (RDF), which is a Web mark-up language that provides basic ontological primitives, has been developed by the W3C (Beckett, 2004). RDF is a language for representing meta-information about resources in the World Wide Web. However, by generalising the concept of a “Web resource”, RDF can also be used to represent information about things that can be identified on the Web, by means of URIs. The DARPA Agent Markup Language + Ontology Inference Layer (DAML+OIL) extends RDF with a much richer set of modelling primitives (Rapoza, 2000). The DAML+OIL have been submitted to W3C as a starting point for the Web Ontology Working Group and led to the creation...
Ontological Engineering in Pervasive Computing Environments

of the standard for Web ontologies, namely the Web Ontology Language (OWL) (McGuinness, 2004; Noy et al., 2001).

In this chapter, we survey the research that has been carried out regarding ontology and knowledge engineering and try to map the key findings to the requirements of a PCE. We present standards and tools that are being used for the development of ontologies. Furthermore, we discuss efforts regarding the management of multiple, overlapping, and evolving ontologies and the semantic mappings among them. We also depict how ontologies can help in searching mechanisms and in profiling users and data sources. Moreover, due to the diverse nature of PCE using interoperability standards is not a sufficiently scalable approach. Hence, we show the important role that ontologies can play in achieving service interoperability, a key requirement in PCE. Finally, we present the efforts that have been carried out linking the pervasive computing paradigm with ontologies and illustrate with an example our vision that brings these two worlds together in terms of semantics usage.

The rest of this chapter is structured as follows: In section 2, we outline our vision of a service oriented PCE and some arguments in the use of the ontology paradigm. Section 3 deals with ontological engineering aspects by providing definitions related to ontologies, presenting current Semantic Web standards for developing ontologies and describing the work that has been carried out regarding their management and manipulation. Section 4 introduces issues related to how ontologies might be used to provide semantic searching mechanisms related to information retrieval. In this section also, we give an overview of ontology-based profiling. In section 5, we portray how ontologies assist in providing semantic based service interoperability and how ontologies may be useful in a PCE. Section 6 outlines some several examples of semantic-based service interoperability through ontologies. Finally in section 7, we provide our conclusions and outline future research directions.

A SERVICE-ORIENTED PERVERSIVE COMPUTING ENVIRONMENT

Service-oriented architectures focus on application-level interoperability, in terms of well-established service ontologies (Martin, 2003; Fensel, 2002). This is achieved by means of well-defined interfaces to various software components. In this fashion, rapid integration of existing functionality is achieved in the design and implementation of new applications. Object Management Group’s (OMG) Common Object Request Broker Architecture (CORBA) (Orfali, 1998) is an example of such widespread service-oriented architecture. One of the most important issues in a PCE is the establishment of ad-hoc relationships between applications or between applications and devices. As a result, the use of a common mechanism, which will be able to provide interoperability in a dynamic fashion, is essential. Furthermore, the separation of the application’s functionality from the adopted communication schemes is of considerable importance, as it enables system developers to distinguish between service functionality (e.g., functional and no-functional context-aware service features, as referred in McIlraith et al., 2001) and system functionality. Hence, the deployment of a PCE reduces to the problem of ad-hoc service discovery, composition (e.g., WSFL, service composition context-aware paradigm) and execution with minimum a-priori knowledge of the service’s functionality. While OWL-S is based on the combination of OWL, WSDL and SOAP, WSMO uses F-Logic and XML-based features of Web Services (WS). Based on the previous languages, significant research has been devoted to semantic WS (e.g., ODE SWS [Gómez et al., 2004], METEOR [Aggarwa et al., 2004]).

However, interoperability is a difficult task to accomplish in a highly heterogeneous and volatile environment. Therefore, for service-oriented architectures to be applicable a well-defined mechanism is required to describe the communication schemes which are employed, provide data with
semantics and describe the profiles and the policies of the entities that constitute the PCE. Furthermore, ontologies may be employed for defining entity-specific policies that govern the context access, usage, and manipulation. They may also provide abstract descriptions of the physical and software components (i.e., conceptual modelling via ontology paradigm) that comprise the PCE. Hence, considerable flexibility in deploying such environments is provided.

Accurate and expressive context models are required to represent context in an efficient and effective way, as the manipulation of the context will be based on its model. The model will determine the variety of the actions that may be applied on it, as well as the accuracy of the results. Moreover, context knowledge and manipulation enable the system to be context aware and proactive and employ more efficient decision making mechanisms. Therefore, we consider the context representation to be of great importance, since formal context modelling (in terms of developing an object-centred or frame-based knowledge manipulation system) achieves multirepresentation knowledge, by rendering heterogeneous nature of PCEs. Moreover, model-mapping and context multirepresentation techniques should be employed to transform the whole context or part of it from one form (i.e., representation formulae or dissimilar semantic conceptual model) to another (e.g., [Ding et al., 2002]). As long as the appropriate mappings exist, the entities may exchange contextual information while preserving their own context model (i.e., localised ontologies [Bouquet, 2002]). Additionally, inference engines (e.g., rule-based systems [Friedman, 2005]) may be used to deduce previously unknown derivative context from primitive contextual information. In this fashion, a pervasive computing system is able to produce knowledge and provide proactivity, assisting the user in an unobstructive way. Last but not least, the profile and policy management for the descriptions of the physical and software entities and the definition of the rules that govern the context access, usage, and manipulation may be based on the ontological engineering paradigm. Therefore, we envisage an environment where all entities interact with services through abstract interfaces and argue that the use of Semantic Web technologies is useful in achieving semantic based service interoperability in PCEs.

ONTLOGICAL ENGINEERING

The word ontology comes from philosophy, where it means a systematic explanation of being. In the last decade, this word has become relevant to the knowledge engineering community. The authors in Guarino (1995) propose the words “Ontology” (with capital “o”) and “ontology” to refer to the philosophical and knowledge engineering concepts, respectively. There are a lot of relevant definitions, but we keep only two terms that are relevant, to some extent, to the pervasive computing paradigm. In Neches et al. (1991) an “ontology” is defined as follows:

An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary.

Such descriptive definition informs us what to do to build an ontology by giving some vague guidelines: This ontology description denotes how to identify basic terms and relations between terms and how to identify rules combining terms and their relationships. Through such a definition one can deduce that ontology does not include only terms that are explicitly defined in, but also the knowledge that can be inferred from it. In this point, the authors in Guarino (1995) consider an ‘ontology’ as:

A logical theory, which gives an explicit, partial account of a conceptualisation, where conceptualisation is basically the idea of the world that a person
or a group of people can have. Consecutively, one may consider that the ontology distinguishes terms that are mainly taxonomies from dissimilar human senses and models a specific domain in a ‘deeper’ way by providing well-documented restrictions on such domain semantics.

In the pervasive computing area it can be claimed that ontology is used to provide a vocabulary with explicitly defined and machine comprehensible meaning. In multi-agent systems, ontologies are essential for providing a common vocabulary and a shared perception of domain knowledge to allow communication in an open environment. An ontology provides an explicit specification of the structure of a certain domain and includes a vocabulary for referring to the subject area, and a set of logical statements expressing the constraints existing in the domain and restricting the interpretation of the vocabulary. Thus, the developing of ontologies, especially for pervasive environments, provides the means by which distributed software components share common semantics of the terms being used for communication and knowledge representation.

**Ontology Languages for the Semantic Web**

As already mentioned, several standards for Web ontologies have emerged. We will provide a brief description of the most widely used standards, namely RDF, DAML+OIL and OWL.

The RDF is a language for representing information about resources in the WWW. It is particularly intended for representing metadata about Web resources, such as the title, author, and modification date of a Web page. However, by generalising the concept of a “Web resource”, RDF can also be used to represent information about things that can be identified on the Web, even when they cannot be directly retrieved on the Web. Examples include information about items available from online shopping facilities (e.g., information about specifications, prices, and availability). RDF is intended for situations in which this information needs to be processed by applications, rather than only being displayed to people. RDF provides a common framework for expressing this contextual information that it can be exchanged between applications without loss of meaning. It represents a simple graph model and uses a well-documented XML schema for datatypes by identifying all elements by means of URIs. The structure of any expression in RDF is a triplet (Figure 1), consisting of a subject, an object, and a predicate (also called property).

RDF provides a way to express simple statements about resources, using named properties and values. However, RDF user communities also need the ability to define the vocabularies (i.e., terms or taxonomies in ontology community) they intend to use in those statements, specifically, to indicate that they are describing specific kinds or classes of resources, and will use specific properties in describing those resources. RDF itself provides no means for defining such application-specific classes and properties. Instead, such classes and properties are described as an RDF vocabulary, using extensions to RDF provided by the RDF Schema (RDFS) (Brickley, 2004). The RDFS provides the facilities needed to describe such classes and properties, and to indicate which classes and properties are to be used together. In other words, RDFS provides a type system for RDF. The RDFS type system is somehow similar

---

**Figure 1. RDF triplet**

```
| Subject | Predicate | Object |
```

```
- rdf:Subject
- rdf:Object
- rdf:Predicate
```

*The SPO conceptual model of RDF*
to the type systems of object-oriented programming languages, such as Java.

DAML+OIL is also a semantic markup language for Web resources. It builds upon earlier W3C standards such as RDF and RDFS, and extends these languages with richer modelling primitives (e.g., in terms of class subsumption or object properties). It adds the familiar ontological primitives of object-oriented and frame-based systems, and the formality of a very expressive description logic (i.e., an ALC Description Logic language).

OWL is a semantic markup language for publishing and sharing ontologies on the World Wide Web. OWL is intended to be used when the information contained in documents needs to be processed by applications, as opposed to situations where the content only needs to be presented to humans. OWL can not only be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms, but also may serve implicitly representing inferred terms or whole taxonomies by means of logical reasoners (i.e., applications, such as RACER, FaCT, DLP (Haarslev, 1999), based on Description Logics axioms that provide inferred knowledge with respect to certain First Order Logic algorithms, such as Structural Subsumption and Tableau algorithms). OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDFS, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web. OWL is an enhanced revision of DAML+OIL Web ontology language. OWL provides three increasingly expressive sublanguages designed for use by specific communities of implementers and ontology developers. The OWL class (i.e., Concepts and conceptual formulas in Description Logics) expressions are built with Knowledge Representation primitives that are properties. These will be organised in two groups. The first group includes primitives defined for OWL Lite while the second group includes primitives defined for OWL DL and OWL Full.

In the previous paragraph, we discussed a set of concept models (i.e., interpretations in the ABox that satisfy the TBox formulas) that are decidable and satisfiable, but with restricted expressiveness due to representation of an sublanguage of the ALC DL. In this context, there is no disjunction and negation concept formulas and the property restrictions (owl:Restriction) of objects (i.e., the instances of concepts in the ABox) are set to cardinality (owl:cardinality).

The latter group stands for fully expressive OWL languages, but the OWL DL is the only sublanguage that can be considered as a machine comprehensible ALCN Description Logic language for the Semantic Web. The OWL Full is not a decidable language in terms of Description Logics, since the ABox and the TBox may overlap (e.g., an instance of the ABox can be interpreted as a concrete concept of the TBox) according to their concept definitions and there is no such a reasoner (i.e., DL algorithm) that can infer whether an interpretation expressed in OWL Full is satisfiable or not. The OWL DL is a DL language enabling the definitions of conjunction, disjunction, and negation for any interpretation of the concept models and deals with certain value (i.e., owl:allValuesFrom, $\forall R.C$ in DL semantics), existential (owl:someValuesFrom, $\exists R.C$ in DL semantics) and number (e.g., owl:maxCardinality, $\geq n$ in DL semantics) restrictions with arbitrarily cardinalities of their fillers (i.e., instances that satisfy Roles in a Knowledge Base ABox).

Due to the dynamic nature of the contextual information in PCEs and the vague, temporal or probabilistic knowledge representation, there have been many extensions in OWL DL language in order to cope with such models (e.g., imprecise information representation, probabilistic knowledge concept definitions, temporal information representation). Such OWL DL languages deal with fuzzy context modelling (e.g., the proposed
temporal contextual concepts (e.g., the proposed [Artale, 1998]) and probabilistic knowledge representation in terms of Bayesian Networks (e.g., the proposed P-OWL language). Such OWL extensions are based on novelty extensions of the classical DLs (e.g., the OWL is a kind of a SHIN language) with the use of Modal Logics (Elgesem, 1997) that provide different world assumptions in well-defined DL Tboxes.

Moreover, in order to deduce new knowledge by reasoning over contextual facts (i.e., set of instances, such as the Abox interpretations), certain extensions in the ontological community by representing rules have been proposed and standardised. The SWRL (Horrocks et al., 2003) is a language for a Semantic Web Rule Language based on a combination of the OWL DL and OWL Lite sublanguages with the Unary/Binary Datalog RuleML sublanguages of the Rule Markup Language (Boley, 2001). Such language extends the set of OWL axioms to include Horn-like clauses. Hence, Horn-like clauses can be combined with an OWL knowledge base by defining rules for the ABox instances reasoning. A high-level abstract syntax is provided that extends the OWL abstract. An extension of the OWL model-theoretic semantics is also given to provide a formal meaning for OWL ontologies including rules written in this abstract syntax.

Managing the Ontology

Taking into account the distributed nature of a PCE, several key challenges emerge concerning the use of ontologies. One has to be able to deal with multiple and distributed ontologies (i.e., distributed knowledge representation in terms of ontological modularity of taxonomies) to enable reuse, maintainability, evolution, and interoperability. Furthermore, support is needed in managing the evolution of multiple and distributed ontologies, in order to ensure consistency. In fact, the ability of logical reasoners to link independent ontology modules, to allow them to be separately maintained, extended, and re-used is one of their most powerful features. In PCE we encounter large and distributed ontologies (i.e., in terms of number of concepts), representing semantically similar contexts (e.g., equivalent concepts or roles).

In highly distributed systems such as the Semantic Web, modularity naturally exists in terms of conceptual modelling. Large ontologies (e.g., medical or biological ontologies such as GALEN (Rector, 2003; Rector et al., 1999) may be split in smaller and reusable modules in order to meet the needs of ontology maintenance and distributed reasoning. Requirements for managing ontologies in such a distributed environment include the following:

- The semantic loose coupling, which refers to the conceptualisation sets of ontologies that are irrelevant (in terms of concept perception).
- The self-standing or self-containment of concepts is the idea of defining concepts that enable the automatic inference of other, classified concepts without having to access contextual information in other ontologies.
- The ontology consistency, which performs integrity checks to assess whether relevant knowledge in other systems has changed, and, updates the self-standing ontological module, if needed.

In Maedche (2003), the authors present a framework for managing multiple and distributed ontologies in a Semantic Web environment. They describe a conceptual modelling framework for single, multiple, and distributed ontologies, which renders feasible the realisation of ontology-based systems using well-established technologies such as relational databases. The authors make use of the object-oriented paradigm and extend it with simple deductive features. In their approach, information is organised in Ontology-Instance
models, which contain both ontology entities and instances. They introduce the “single ontology consistency” and the “dependent ontology consistency” definitions and propose an evolution strategy that ensures the consistency of the evolving ontologies by unambiguously defining the way in which elementary changes will be resolved.

In another approach, discussed in Rector (2003), the key is modularity. Modularity is acknowledged as a key requirement for large ontologies in order to achieve reuse, maintainability, and evolution. Rector, motivated by the fact that mechanisms for normalisation are standard for databases, proposes the concept of normalisation for ontologies. He proposes a two-step normalisation implemented using OWL or related description logic-based formalisms. For the first step, namely the “ontological normalisation”, he makes use of the analysis in Guarino (2000). For the second step (implementation normalisation) he proposes an approach based on decomposing the ontology into independent disjoint skeleton taxonomies restricted to simple trees, which can, then, be recombined using definitions and axioms to explicitly represent the relationships between them. The main requirement, though, for implementation normalisation is that the modules to be reused can be identified and separated from the whole, so as to evolve independently.

In Guarino (1999), the authors introduce the OntoSeek, a system that employs an ontology server for ontology management. The server provides an interface for applications willing to access or manipulate an ontology data model (i.e., a generic graph data structure), and facilities maintaining a persistent Lexical Conceptual Graph (LCG) database. End users and resource encoders can access the server to update the LCG database encoded in a markup language, (e.g., XML). The main function of the ontology server is to create, edit, evaluate, publish, maintain, and reuse ontologies. Particularly important is the ability to support the collaborative works through the Web.

Semantic Mappings Among Ontologies

Clearly, it would be desirable in a PCE that the various services are capable of interoperating with existing data sources and consumers. It is very difficult to build a consensus about the terminologies and structures that should be used. In order to introduce the significance of the ontology matching issue in the computing paradigm, we support the Semantic Web to make information on the World Wide Web more accessible using machine-readable metadata. Since the ontological development deals with conceptualising specific domains, the notion of context modelling could be imperative by using specific domain ontologies. Thus, an ontology is contextualised when its concepts are kept local (i.e., self-standing concepts that are not shared with other ontologies) and are associated only with the concepts of other ontologies, through well-defined explicit mappings (or implicit mappings, through distributed reasoning tasks, such as Compilation of Implied Knowledge with Queries).

An architecture is needed for building a Web of data by allowing incremental addition of sources, where each new source maps to the sources considered most convenient. Therefore, information processing across ontologies is only possible when semantic mappings between the ontologies exist. Technological infrastructure for semantic interoperability between semantically autonomous communities (i.e., ontological development that is compliant with the requirements in subsection 3.2) must be based on the capability of representing local ontologies and contextual mappings between them, rather than on the attempt to create a global and possibly shared conceptualisation. Manually finding such contextual mappings is tedious, error prone, and clearly not possible on a large scale. This is what researchers call semantic heterogeneity in the Semantic Web paradigm, namely, a situation in which different meanings use the same concept
Ontological Engineering in Pervasive Computing Environments

to mean the different things and use different granularity to describe the same domain forming explicit different perspectives of the world. In this context, the lack of uniformity prompts for the meaning negotiation among different agents, since they do not understand each others as they use languages with heterogeneous semantics.

A lot of work has been performed during the last years in this direction, which resulted in the development of systems that provide mappings semi-automatically (Doan, 2003; Halevy, 2003; Fowler et al., 1999; Chalupsky, 2000). In this way, given the semantic mappings between different ontologies, one may match them in order to be able to reason about a specific domain.

Given two ontologies, the ontology-matching problem is to find semantic mappings between them. The simplest type of mapping is the one-to-one mapping between the elements of the different ontologies, namely each element of one ontology is associated with at most one element of the second ontology and vice versa. Mappings between different types of elements are possible, meaning that a property may be mapped to a class. More complex mappings are possible, such that a class or an instance may map to a union of classes or instances. In general, a mapping could be specified as a query that transforms instances, classes and properties from one ontology into instances, classes, and properties of another.

To illustrate the above, suppose that we want to provide mappings between two WWW sites that sell books. In Figure 2, the ontology-matching problem between an author-centric ontology and a book-centric ontology is depicted. A matchmaking process must provide semantic mappings between the nodes in the two ontologies in an automatic manner.

In Doan (2003), the authors describe GLUE, a system that employs machine-learning techniques to determine such mappings. Given two ontologies, for each concept in one ontology the system finds the most similar concept in the other ontology. A key feature of the GLUE is that it uses multiple learning strategies, each of which exploits well a different type of information either

---

**Figure 2. Ontology matching problem**

---

![Ontology matching problem diagram](image-url)
in the data instances or in the taxonomic structure of the ontologies.

The first issue to address such conceptual matching is the notion of similarity between two concepts. The approach in Doan (2003) is based on the observation that many practical measures of similarity can be defined based solely on the joint probability distribution of the concepts involved. The second challenge is that of computing the joint distribution of any two given concepts A and B. Specifically, for any two concepts A and B, the joint distribution consists of $p(a \in A \land B)$, $p(a \in A \land \neg B)$, $p(a \in \neg A \land B)$, $p(a \in \neg A \land \neg B)$, where a term such as $p(a \in A \land \neg B)$, is the probability that an instance in the domain belongs to concept A but not to concept B. GLUE, also, uses a multi-strategy learning approach employing a set of learners and then combining their predictions using a meta-learner. Furthermore, it exploits domain specific constraints using general heuristic algorithms to improve the matching accuracy.

In another approach, described in Halevy (2003), the authors provide a mediating schema between RDF and XML data sources. Most of these sources provide XML representations of the data and not RDF; XML and the applications that depend on it rely not only on the domain structure of the data, but also on its document structure. In order to provide interoperability between such sources, one must map between both their domain structures and their document structures. Second, data management practitioners often prefer to exchange data through local point-to-point data translations, rather than mapping to common mediated schemas or ontologies. Therefore, the authors built the Piazza system, which addresses these challenges. Piazza offers a language for mediating between data sources on the Semantic Web, which maps both the domain structure and document structure. Piazza also enables interoperation of XML data with RDF data that is accompanied by rich OWL ontologies. They proposed a language for mediating between nodes that allows mapping simple forms of domain structure and rich document structure. The language is based on Xquery (Fernández et al. 2005), the emerging standard for querying XML. They also showed that this language could map between nodes containing RDF data and nodes containing XML data.

In Fowler et al. (1999), the authors propose InfoSleuth, a system that can support construction of complex ontologies from smaller component ontologies so that tools tailored for one component ontology can be used in many application domains. Examples of reused ontologies include units of measure, chemistry knowledge, and geographic metadata. The mapping is explicitly specified among these ontologies as relationships between terms in one ontology and related terms in other ontologies.

In another approach, the OntoMorph system of the Information Sciences Institute (ISI) of the University of Southern California aims to facilitate ontology merging and the rapid generation of knowledge base translators (Chalupsky, 2000). It combines two powerful mechanisms to describe Knowledge Base transformations. The first of these mechanisms is syntactic rewriting via pattern-directed rewrite rules that allow the concise specification of sentence-level transformations based on pattern matching, and the second mechanism involves semantic rewriting which modulates syntactic rewriting via semantic models and logical inference. The integration of ontologies can be based on any mixture of syntactic and semantic criteria.

**ONTOLOGY-BASED PROFILING AND INFORMATION RETRIEVAL**

In a typical PCE there exists a vast amount of contextual information available for retrieval. The traditional solution to the problem of information retrieval employs keyword-based searching techniques. This solution is inadequate and, therefore, new searching mechanisms must be developed.
The context should be modelled in a semantically enriched manner, namely by means of ontologies, and searching techniques that make use of this semantic information must be applied in order to provide superior results. Furthermore, PCEs usually integrate sensor networks, which differ considerably from current networked and embedded systems. They combine the large scale and distributed nature of networked systems with the extreme energy constraints and physically coupled nature of embedded control systems. Sensors, however, produce raw data. Ontologies could be used to transform the raw data to a semantically enriched representation that could be queried by an advanced searching mechanism.

The proliferation of Semantic Web technologies, such as ontologies, has led to the development of more sophisticated semantic query mechanisms. Semantic querying techniques will exploit the semantics of content to provide results that are superior to those delivered by contemporary techniques, which rely mostly on the lexical and structural properties of a document. There are a number of proposals for querying data (e.g., Anyanwu et al., 2003) expressed by means of ontologies, but most of them are constrained in such a way that one is able to form queries only of the type “Get all elements that are related to element A through a relationship R” where R is, typically, specified as a path expression. Recently, research is being conducted in order to be able to express queries of the type “How is element A related to element B?”

In Khan (2004), the authors propose a framework for querying databases of a specific context using an ontology-based index structure. The traditional solution employs keyword-based search. The only documents retrieved are those containing user specified keywords. But many documents convey desired semantic information without containing these keywords. One can overcome this problem by indexing documents according to context and meaning rather than keywords, although this requires a method of converting words to meanings and the creation of a meaning-based index structure. In the discussed work, the authors have solved the problem of an index structure through the design and implementation of a concept-based model using domain dependent ontologies. In other words, the ontology provides index terms/concepts that can be used to match with user requests. Furthermore, the generation of a database query takes place after the keywords in the user request are matched to concepts in the ontology. To achieve this, they employ a proprietary method for query expansion and SQL query generation.

In Wang et al. (2003), a system for classifying text documents based on ontologies for representing the domain specific concept hierarchy (an example is depicted in Figure 3) is presented. Many researchers have shown that similarity-based classification algorithms, such as k-nearest neighbour and centroid-based classification, are very effective for large document collections. However, these effective classification algorithms still suffer disadvantages from high dimensionality that greatly limit their practical performance. Finding the nearest neighbours in high dimensional space is very difficult because the majority of points in high dimensional space are almost equi-distant from all other points. The k-nearest neighbour classifier is an instance-based classifier. This means that a training dataset of high quality is particularly important. An ideal training document set for each particular category will cover all the important terms, and their possible distribution in this category. With such a training set, a classifier can find the true distribution model of the target domain. Otherwise, a text that uses only some keywords out of a training set may be assigned to the wrong category. In practice, however, establishing such a training set is usually infeasible. Actually, a perfect training set can never be expected. By searching the concept hierarchy defined by a domain specific ontology, a more precise distribution model for a predefined classification task can be determined.
The experiments that were carried out indicated that, by using this approach, the size of the feature sets can be effectively reduced and the accuracy of the classifiers can be increased.

An extension of the ontological reasoning, in terms of information retrieval, is discussed in Maedche et al. (2001). In this work, the authors, in order to define and exploit the semantics of domain ontologies, perform the context retrieval process by clustering the pre-classified concepts with several similarity measures. The metrics of the semantic distances among concepts, which are referred to as similarity measures, are the Taxonomy similarity, the Relation similarity, and the Attribute similarity. Such measures form the input to the hierarchical clustering algorithm. In this context, one may deduce about new relations among concepts, something that enriches the notion of the distributed domain ontologies in PCE. Furthermore, the classification process, with respect to the context retrieval, is not feasible with undefined relations among conceptual formulas (i.e., primitive concept, defined concepts, and complex concept definitions).

Moreover, such new relations may denote new activities (i.e., constraints), behaviours (i.e., axioms), and roles on entities or concepts in ontologies. In the pervasive computing paradigm we encounter a large number of interacting entities. The rules that govern that interaction are complex and should be based, among others, on the profiles of entities. According to the Cambridge Dictionary (1980), a profile is “a portrait, drawing, diagram, etc., of a side view of the subject”. A model of an entity’s characteristics, activity, behaviour and its role in the PCE is the entity’s profile. Based on this knowledge, complex mechanisms manage interaction of entities in the most efficient and effective way.

Several methods have been proposed for using ontologies for user and data sources profiling, as well as for describing communication mechanisms. Hence, complex queries over data sources may be exploited and auto-learning systems may propose content sources to users based on their preferences.

In Middleton et al. (2004) the authors explore an ontological approach to user profiling in the context of a recommender system. Their
system, called Foxtrot, addresses the problem of recommending online research papers to academic researchers. Their ontological approach to user profiling is a novel idea to recommender systems. Researchers need to be able to search the system for specific research papers and have interesting papers recommended. Foxtrot uses a research-paper-topic ontology to represent the research interests of its users. A class is defined for each research topic and is-a relationships are defined where appropriate. Their ontology is based on the CORA (McCallum et al., 2000) digital library, as it classifies computer science topics and has example papers for each class. Research papers are represented as term vectors, with term frequency/total number of terms used for a terms weight; terms represent single words in a paper’s text. Research papers in the central database are classified by an IBk (Kibler, 1991) classifier, a k-nearest neighbour type classifier, which is boosted by the AdaBoostM1 algorithm (Freund, 1996).

In Parkhomenko et al. (2003), the authors present an approach that uses ontologies to set up a peer profile containing all the data, necessary for peer-to-peer (P2P) interoperability (Milojicic et al., 2002). The use of this profile can help address important issues present in current P2P networks, such as security, resource aggregation, group management, etc. They also consider applications of peer profiling for Semantic Web built on P2P networks, such as an improved semantic search for resources not explicitly published on the Web but available in a P2P system. They developed the ontology-based peer profile in RDF format and demonstrated its manifold benefits for peer communication and knowledge discovery in both P2P networks and Semantic Web.

In another approach regarding P2P networks, the authors in (Nejdl et al., 2003), focus on peer profiling so as to enable complex user queries and routing algorithms. RDF-based P2P networks allow complex and extendable descriptions of resources, and they provide complex query fa-

**SEMANTIC-BASED SERVICE INTEROPERABILITY THROUGH ONTOLOGIES IN PERSVASE ENVIRONMENTS**

Although, the Web was once just a content repository, it is, now, evolving into a provider of services. Web-accessible programs, databases, sensors, and a variety of other physical devices realise such services. In the next decades, computers will most likely be ubiquitous and most devices will have some sort of computing functionality. Furthermore, the proliferation of intranets, ad-hoc, and mobile networks sharpen the need for service interoperation. However, the problem of service interoperability arises because today’s Web is designed primarily for human use. Nevertheless, an increased automation of services interoperation, primarily in business-to-business and e-commerce applications, is being noted. Generally, such interoperation is realised through proprietary APIs that incorporate hard-coded functionality in order to retrieve information from Web data sources. Ontologies can prove very helpful in the direction of service description for automatic service discovery, composition, interoperation, and execution.

In McIlraith et al. (2001), the authors present an agent technology based on reusable generic procedures and customising user constraints that exploits and showcases WS markup. To realise their vision of Semantic WS, they cre-
Ontological Engineering in Pervasive Computing Environments

ated semantic markup of WS that makes them machine understandable and easy to use. They also developed an agent technology that exploits this semantic markup to support automated WS composition and interoperability. Driving the development of their markup and agent technology are the automation tasks that semantic markup of WS will enable, in particular, service discovery, execution, and composition and interoperation. Automatic WS discovery involves automatically locating WS that provide a particular service and adhere to requested properties. Automatic WS execution involves a computer program or agent automatically executing an identified WS. Automatic WS composition and interoperation involves the automatic selection, composition, and interoperation of appropriate WS to perform some task, given a high-level description of the task’s objective. They use ontologies to encode the classes and subclasses of concepts and relations pertaining to services and user constraints. The use of ontologies enables sharing common concepts, specialisation of these concepts, and vocabulary for reuse across multiple applications, mapping of concepts between different ontologies, and the composition of new concepts from multiple ontologies.

In Gibbins et al. (2003), the authors describe the design and implementation of an ontologically enriched WS system for situational awareness. Specifically, they discuss the merits of using techniques from the multi-agent systems community for separating the intentional force of messages from their content and the implementation of these techniques within the DAML Services model. They identify that the world of WS may be characterised as a world of heterogeneous and loosely coupled distributed systems where adaptability to ad-hoc changes in the services offered by system components is considered advantageous. The term loosely coupled, means that the interactions between system components are not rigidly specified at design time. Instead, system components may opportunistically make use of new services that become available during their lifetime without having been explicitly told of their existence from the outset. The task of searching for a system component which can perform some given service (i.e., service discovery) is the enabling technique that makes loosely-coupled systems possible, and provides a process by which system components may find out about new services being offered. Service descriptions are more complex expressions, which are based on terms from agreed vocabularies and attempt to describe the meaning of the service, rather than simply ascribing a name to it. An essential requirement to service discovery is service description. A key component in the semantics-rich approach is the ontology. In the conventional WS approach exemplified by Web Services Definition Language (WSDL) or even by DAML Services, the communicative intent of a message (e.g., whether it is a request or an assertion) is not separated from the application domain. This is quite different from the convention from the Multi-Agent Systems world, where there is a clear separation between the intent of a message, which is expressed using an Agent Communication Language (ACL), and the application domain of the message, which is expressed in the content of the message by means of domain-specific ontologies. In this approach, there is a definite separation of the intent of the messages and the application domain.

In Medjahed et al. (2003), the authors propose an ontology-based framework for the automatic composition of WS. An important issue in the automatic composition of WS is whether those services are composable (Berners-Lee, 2001). Composability refers to the process of checking if WS to be composed can actually interact with each other. A composability model is proposed for comparing syntactic and semantic features of WS (Figure 4). The second issue is the automatic generation of composite services. A technique is proposed to generate composite service descriptions while obeying the aforementioned composability rules. Such technique uses as
input a high-level specification of the desired composition. This specification contains the list of operations to be performed through composition without referring to any composite service. Based on their composability model, they propose an approach for the automatic composition of WS. This approach consists of four conceptually separate phases: specification, matchmaking, selection, and generation. The authors define an XML language, which adopts an ontology-based model to cater for Semantic Web-enabled WS, for the specification of composite services. Once specifications are provided, the next step is to generate appropriate composition plans using a matchmaking algorithm. The general principle of the matchmaking algorithm is to map each operation of the composite service to one or more operations of existing services. At the end of the matchmaking phase, several composition plans may have been generated. To facilitate the selection of relevant plans, they define three parameters for the quality of composition: ranking, relevance, and completeness. The ranking of a composition gives an approximation of its “importance”, the relevance gives an approximation of the composition soundness and the completeness gives the proportion of composite service operations that are composable with other services. The last phase in their approach aims at generating a detailed description of a composite service. This description includes the list of outsourced services, mappings between composite services and component service operations, mappings between messages and parameters and flow of control and data between component services.

Let us imagine a PCE where all devices interoperate through services. In this way, a service-oriented architecture for functionality integration is being employed. Therefore, the problem of deploying a PCE reduces to the problem of service description and ad-hoc service discovery, composition, and execution.

We propose that the services/agents that represent the various devices share a common communication scheme, possibly by means of an ACL or a common ontology that describes the communication details. However, they do not share any a priori knowledge about the se-
Ontological Engineering in Pervasive Computing Environments

The semantics of the functionality that these devices provide. As described above, through the use of the device-specific ontologies one can achieve runtime service discovery and execution based on an abstract description of his objectives. In other words, a device will query its environment, giving the description of its objectives and employing complex semantic querying mechanisms. After obtaining a service reference, it will proceed with its execution. As new devices join the PCE and while others abandon it, the cycle of service description, discovery, and execution will continue forever with minimum external interference.

Furthermore, the work that has been carried out in the field of ontology management and ontology mapping ensures that the consistency of the ontologies is maintained as they evolve throughout time. Furthermore, webs of complex services are being created as devices join in or withdraw from the PCE and as new composite services are being created based on noncomposite (i.e., primitive) or even other composite services. This is only possible when one can provide semantic context mappings between the different services and automatically produce semantic descriptions of the composite services. Hence, a complicated hierarchy of services may be created. Such a hierarchy should be both manageable and highly flexible.

EXAMPLES OF SEMANTIC-BASED SERVICE INTEROPERABILITY THROUGH ONTOLOGIES IN PCE

To illustrate the key issues, consider the following simple scenario (Figure 5). Bob has an important professional meeting (i.e., human society context) in his office (OfficeA) with Alice. For this meeting it is important that Bob has a certain postscript file (FileA) printed (i.e., application context). Unfortunately, the printer in his office (PrinterA) is not capable of printing postscript files. However, there is one such printer (PrinterB) in another office (OfficeB), but Bob is unaware of its existence (i.e., service registry context). In addition, the two offices are adjacent (i.e., spatial context). Bob is engaged in the activity of editing (i.e., human activity-oriented context) the file and finishes just 15 minutes before his appointment (e.g., temporal context). The PCE, by recording the time at which Bob finishes editing the file and knowing that the appointment is scheduled in 15 minutes, (i.e., contextual history) infers that Bob wants this file for the appointment (e.g., applying deduction rules over certain contexts). Furthermore, it judges that there is sufficient time for Bob to go to the next office and print the file and, consequently, issues the suggestion (i.e., ambient notification to human).

There are several types of semantically enriched contexts in this scenario. First of all the system knows the location of Bob, as well as the locations of the offices within the building. It also has knowledge about the equipment that is present at the offices and its capabilities (i.e., terminal profile). The system has access to his personal agenda and monitors the activities that he is engaged in (i.e., user profile). It is, also, capable of associating any activity with specific entities, which, in turn have specific attributes (i.e., inference over context facts). Furthermore, it employs an inference engine to deduce new contextual information (Bob wants the files for his appointment) and uses this information to issue suggestions.

In another scenario, a system for observing and assisting elderly people, who are alone at home, may be considered. Ontologies may be used to define simple cases (e.g., the person is sleeping for the past fifteen hours or has been in the bathroom for too long) in which the system must take specific actions, wake the patient by ringing a bell or informing the attending doctor. Furthermore, with the addition of a rule engine, complex conditions may be defined that were, previously, required to be diagnosed by a trained doctor. For instance, the system may constantly
monitor the patient’s Electro Cardiogram (ECG) and, based on time depended composite rules, may decide that the patient requires immediate medical attention.

Research has been reported in the association of Semantic Web technologies and pervasive computing. In Chen et al. (2004), the authors describe a framework for an agent-based PCE. Central to their framework is the presence of an intelligent context broker that accepts context related information from devices and agents in the environment as well as from other sources, including information available on the Web describing the space and events and activities scheduled for it. The context broker integrates and reasons over this information to maintain a coherent model of the space, the devices, agents, and people in it, and their associated services and activities. A key to the realisation of this architecture is the use of a set of common ontologies (e.g., specialising an Upper Level Ontology or semantics similarity measures). They identify ontologies as a key component for building open and dynamically distributed PCEs in which agents and devices share contextual information. They describe the use of OWL and other tools for building the foundation ontology for the Context Broker Architecture (CoBrA). The CoBrA ontology models the basic concepts of people, agents, place, and presentation events in an intelligent meeting room environment. It provides a vocabulary of terms for classes and properties suitable for building practical systems that model context in PCEs.

In Wang (2003), the author describes “The context broker”, a broker-centric agent architecture, which aims to support context-aware computing in intelligent spaces. In this architecture, a broker maintains and manages a shared context model on the behalf of a community of devices and agents. It provides necessary common services including ontology mediation and teamwork support, and enforces policies for security, trust, and privacy. A reasoning engine is responsible for reasoning with ontological and contextual knowledge. As ontological knowledge, the author identifies the static knowledge derived from the underlying
ontology model (expressed in OWL), and as contextual knowledge, the dynamic knowledge that is inferred from acquired situational information.

In Ranganathan et al. (2003) and Ranganathan-GAIA (2003), the authors describe GAIA, an infrastructure for Smart Spaces, which are ubiquitous computing environments that encompass physical spaces. GAIA converts physical spaces and the ubiquitous computing devices they contain into a programmable computing system. It offers services to manage and program a space and its associated state. GAIA is similar to traditional operating systems in that it manages the tasks common to all applications built for physical spaces. Each space is self-contained, but may interact with other spaces or modular ontologies (e.g., via context mappings). GAIA provides core services, including events, entity presence (e.g., devices, users and services), discovery, and naming. By specifying well-defined interfaces to services, applications may be built in a generic way so that they are able to run in arbitrary active spaces. The core services are started through a bootstrap protocol that starts the GAIA infrastructure. Finally, GAIA allows application developers to specify different behaviours of their applications for different contexts. The use of contextualised ontologies makes easier for developers to specify context-sensitive behaviour.

One of the main uses of ontologies is that it allows developers to define all the terms that can be used in the environment. Ontological engineering allows the attachment of precise semantics to various terms and the clear definition of the relationships between different terms. Hence, it prevents semantic ambiguities where different entities in the environment have different ideas from what a particular term means. Different entities in the environment can refer to the ontology to get a definition of a term, in case they are not sure (e.g., due to the imprecise conceptual modelling or the nature of the contextual information). Furthermore, developers employ ontologies for describing both entities (e.g., concepts) and contextual information (domain specific conceptual modelling). In this manner, they can determine whether these descriptions are valid with respect to the schema defined by the ontology. When a new entity is introduced into the system, its description can be checked against the existing ontology to see whether it is satisfiable (Roman et al., 2002). If the description is not consistent with the concepts described in the ontology, then the description is faulty. Moreover, ontologies are used to provide semantic discovery of objects and services. Conventional object registries provide a limited capability for object discovery and the existing discovery protocols support limited ability to spontaneously discover entities on a network. However, in PCEs it must be possible to discover only the “relevant” entities, without knowing, in advance, what will be relevant (Amann et al., 2003). This process has also been termed “match-making”. Finally, a key feature of applications in PCEs is that they are context-aware, i.e., they are able to obtain the current context and adapt their behaviour to different situations.

CONCLUSION AND FUTURE WORK

The vision of pervasive computing introduces structural issues, as in a typical ubiquitous environment there will be numerous devices with computing capabilities that need to interoperate. Therefore, we cannot rely on shared a priori knowledge based on commonly accepted standards to resolve such issues. Software components must adapt to their environment at runtime to integrate their functionality with other software components seamlessly. We argued that the concept of Semantic Web, by the means of ontologies, could prove to be a satisfactory solution to the aforementioned problem, as it enables the system and the software providers to deal with semantically enriched specifications of services and semantically enriched data. “Models” based on ontologies may be used to provide objects, and
mainly contextual information, with semantics. The context representation is thought to be of the utmost importance, as context knowledge and manipulation enables a system to be context aware and proactive and to employ more effective and efficient decision-making mechanisms. In addition, profile and policy management could be based on ontology models, in order to provide maximum flexibility for describing entities and rules regarding resource usage and manipulation.

Moreover, we surveyed the research that has been carried out regarding ontologies and tried to map the key findings to the requirements of a PCE. We have discussed efforts dealing with the management of multiple, overlapping, and evolving ontologies and the semantic mappings among them. We, also, depicted how ontologies can help in searching mechanisms and in profiling of users and data sources. Furthermore, due to the diverse nature of PCEs, the use of interoperability standards is not a sufficiently scalable approach. Hence, we showed the important role that ontologies could play in achieving service interoperability, a key requirement in PCEs, and depicted how one may use ontologies in the deployment of PCE.

Research on the association of ontologies and PCEs is rather restricted. Mainly the artificial intelligence community is carrying out the research regarding ontologies, whilst the research regarding pervasive computing is mostly conducted by the network community. Thus, the need of collaboration between the two communities is clearly visible. Nevertheless, a significant trend is identified. More and more researchers acknowledge the necessity of incorporating ontologies into the development of systems for supporting PCE. In this way, the emerging standard for Web ontologies, such as the (OWL), may prove to be a significant factor in the road ahead.

However, there are several future research directions that should be considered. Description logics may perform well in defining domain vocabularies and concepts but lacks in describing modal entities, such as temporal entities, vague knowledge, and probabilistic knowledge. Several extensions have been proposed (Baader et al., 2003) in that direction. These extensions include the following: the ability to predefined specific predicates over concrete domains (e.g., the greater-than or less-than operators over the domain of integers), the use of modal extensions in order to define concepts like belief and obligation, temporal extensions for describing time dependence, representation of uncertain and vague knowledge (probabilistic and fuzzy logic extensions) and nonmonotonic reasoning. These extensions should be incorporated into the Semantic Web standards in order to achieve maximum expressiveness. Last but not least, collaboration between various scientific communities should be employed, so as to see the Semantic Web reach its full potential, pervading in everyday life.

REFERENCES


O’Sullivan, D., & Lewis, D. (2003). Semantically driven service interoperability for pervasive com-
Ontological Engineering in Pervasive Computing Environments


Rector, A., & Rogers, J. (1999). Ontological issues in using a description logic to represent medical concepts: Experience from GALEN. Appeared at IMIA WG6 Workshop: Terminology and Natural Language in Medicine, Phoenix, AZ.


This work was previously published in Web Semantics and Ontology, edited by D. Taniar, pp. 334-363, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 3.8
MASACAD: A Multi-Agent System for Academic Advising

Mohamed Salah Hamdi
University of Qatar, Qatar

ABSTRACT
The evolution of the Internet into the Global Information Infrastructure has led to an explosion in the amount of available information. Realizing the vision of distributed knowledge access in this scenario and its future evolution will need tools to customize the information space. In this article we present MASACAD, a multi-agent system that learns to advise students and discuss important problems in relationship to information customization systems and smooth the way for possible solutions. The main idea is to approach information customization using a multi-agent paradigm.

INTRODUCTION
In our previous work (Hamdi, 2005), we have presented an e-learning system that provides a service to a student that checks whether lecturers are offering information that matches the profile of the student and informs the student of the information found. The student is registered in many courses and seeks course materials from the sites of the different lecturers. These diverse resources can be physically distributed. They also are dynamic so that course materials can be added, updated, or deleted. The student profile, that includes the courses currently attended and possibly much more information, changes also over time because the student can leave a course or register in a new one. All of this means that the customized presentation of information for the student should be updated continuously as new information becomes available. This happens with no user intervention using an autonomous multi-agent system.

In this article, following the same long-term objective of providing a complete e-learning environment for students and aiming for the more general goal of information customization, we describe MASACAD (Multi-Agent System for
ACademic ADvising; “MASACAD” is also the Arabic word for “courses”), a multi-agent system that advises students by adopting a machine learning paradigm. The decision for employing the multi-agent approach combined with machine learning is supported by the following general arguments (more specific arguments are discussed in Section 3.5).

Humans often deal with problem complexity by organizing teams and cooperating in the search for solutions. Group problem solving techniques have evolved into powerful strategies in areas where any single individual would simply be overwhelmed. AI research has paralleled this approach by developing the distributed problem-solving paradigm. Problem-solving systems can distribute among them the processes needed to accomplish a given task.

Given the amount of problems that can be approached through distributed processing, AI has directed significant effort towards exploring possibilities to develop specialized problem-solving systems that can interact in their search for a solution. One way to embody this approach is represented by the multi-agent system paradigm.

An agent is a collection of knowledge and methods that are intended to embody a well-defined functionality. Agents cooperate in solving a problem by contributing various parts of the problem-solving process. Agents can be modified independently and they are well focused on their tasks. Such subsystems are often easier to design, develop, and validate, than their counterparts that combine the equivalent functionality within one system.

Because it is not possible to account for all the aspects of multi-agent problem-solving at development time, a good problem-solver has to be able to compensate autonomously for the dynamic and unpredictable aspects of its execution environment through adaptation. Adaptation is reached by letting the system learn from experience.

In the following, we start with presenting some background material and then introduce the problem of academic advising and argue that a multi-agent system would be a good approach for addressing this problem. The paper then presents the architecture of the proposed system and discusses in detail its individual components. The following section discusses the student evaluation of the system and provides preliminary evidence that it is helpful. Benefits and limitations of MASACAD, future directions of work, and related research are discussed respectively in each of the next sections. The final section presents a short summary of the paper.

BACKGROUND

The MASACAD system makes use of a wide range of technologies. These are discussed briefly in the following paragraphs.

E-Learning: Several years ago, online education was considered as an experimental approach with more disadvantages than advantages. However, today it should be considered not only a complementary educational resource but also a serious alternative that competes with conventional and now classical methods. The adaptation to the new features and services of the e-learning environment is not immediate and requires experience, time, investment, pedagogical and technical resources, and government or campus administration support. At the UAE University there exists enormous interest in the area of online education. Rigorous steps are taken towards the creation of the technological infrastructure (hardware, software, and communications) and the academic infrastructure (course materials, teacher-student communication) for the improvement of teaching and learning. MASACAD, the academic advising system described in this article is to be understood as a tool that uses network technology to support learning and as part of the e-learning environment at the university. We use it to demonstrate the capability of exploiting the digital infrastructure, enabled by the online mode
of teaching and learning, to extract and infer useful knowledge and to point out the potential of computational intelligence in the area of intelligent Web-based education.

**Information Customization:** Building software that can interact with the range and diversity of the online resources is a challenge and the promise of information customization (IC) systems is becoming highly attractive. It would be wonderful if an IC system can identify and present the information with little or no user intervention. The system should also be able to update the presentation as new information becomes available. This will release the user from continually observing the resources. This raises, of course, questions about robustness and persistence of the system. An important point in relationship with IC systems concerns therefore their evaluation. The best test of an IC system’s ability to help a user is a user trial. In the present work, we evaluate the IC system for academic advising both by standard tests (performance of the learning algorithm) and by user trials (performance of the whole system).

**User Modeling:** An IC system is a kind of software that acts in accordance with a user’s preferences in an environment. To realize an IC system acting in accordance with a user’s preferences, user modeling is needed. User profiles are of great importance for information extraction and information customization since they are essential for deciding what kind of information is needed, where this information can be found, how this information can be retrieved, and how this information should be presented to the user. User profiles will therefore have a great influence on the solution to be adopted for implementing an IC system. In our case they will have a strong impact on the multi-agent system to be created.

**Agent Approach:** Agents can be viewed as a new model for developing software to interact over a network. This view has emerged because of the predominance of networks in the world. Information, knowledge, and electronic resources in general, are distributed across a network and programs and methods are needed to access them and present them in a customized manner. Using agents adds a layer of abstraction that localizes decisions about dealing with local peculiarities of format, knowledge conventions, and so on, and thus helps to understand and manage complexity. Agents therefore should be seen as an abstraction that appears to provide a powerful way of conceptualizing, designing, and implementing a particularly complex class of software systems. Multi-agent systems are systems composed of multiple interacting agents, where each agent is a coarse-grained computational system in its own right. The approach of multi-agent systems seems to be a suitable framework for developing IC systems since many of the properties of IC systems or requirements on these systems coincide with those required on multi-agent systems and on agent-based systems in general. The IC system proposed in this article for dealing with the problem of academic advising adopts the multi-agent paradigm.

**Machine Learning:** Human expertise, needed for solving problems, should be transferred and transformed from some source of knowledge to a program. This transfer is usually accomplished by a series of lengthy and intensive interviews between a knowledge engineer, who is normally a computer specialist, and a domain expert who is able to articulate his expertise to some degree. Unfortunately, the productivity of the interviewing process is typically poor for many reasons (see Jackson, 1999). This has led researchers to look upon the knowledge acquisition phase as “the bottleneck problem” of expert systems applications (Feigenbaum, 1977). This dissatisfaction with the interview method has encouraged some researchers to try to automate the process of knowledge acquisition by means of machine learning methods. Artificial neural networks are a particular method for empirical learning. They have proven to be equal, or superior, to other empirical learning systems over a wide range of
domains, when evaluated in terms of their generalization ability (Shavlik, Mooney, & Towell, 1991). Although the almost complete ignorance of problem-specific theory by empirical learning systems may mean that they do not address important aspects of induction, it is interesting to see in the following study, how domain-specific knowledge about academic advising of students can be employed by a domain free neural network learning algorithm.

**Web Mining:** Current Web mining research aims at creating systems that (semi) automatically tailor the content delivered to the user from a Web site. This is usually done by mining the Web — both the contents, as well as the user’s interaction (Cooley, Mobasher, & Srivastava, 1997). To mine data from the Web is therefore different from mining data from other sources of information. For the problem of academic advising at hand, the Web represents the main source of information. A solution to this problem will therefore mine the Web. The Web mining technique adopted, bases on the multi-agent paradigm combined with machine learning and ideas from user modeling.

**THE PROBLEM OF ACADEMIC ADVISING**

We will illustrate our ideas using MASACAD, an example consisting of an e-learning application. In this application the focus is on the academic advising for students.

**Academic Advising**

The general goal of academic advising is to assist students in developing educational plans which are consistent with academic, career, and life goals and to provide students with information and skills needed to pursue those goals. More specifically, advisors will assist students in the following ways:

- Guide students through the university’s educational requirements;
- Assist in scheduling the most appropriate courses;
- Introduce them to pertinent resources;
- Promote leadership and campus involvement;
- Assist in career development;
- Assist students with the timely completion of their degree; and
- Help students find ways to make their educational experience personally relevant.

**Why is a Software Assistant Needed for Academic Advising?**

In order to help the student, improve the advising process and make it easier, and overcome the many problems that may occur, an intelligent assistant in form of a computer program will be of great interest. Such an intelligent assistant will automate the advising process in the university. It will also simplify the task of faculty, staff, students, and professional advisors and make it possible to save time and effort and prevent mistakes. These benefits are added to the many other advantages of any assistant software that is used to solve problems that ordinarily require human expertise.

**Restriction of the General Goal of Academic Advising**

The goal of academic advising, as stated previously, is too general because many experts are involved and because a huge amount of expertise is needed. Hence, realizing an intelligent software assistant that is able to deal with all the details shown previously will be too difficult, if not impossible. In the following implementation we will therefore restrict academic advising and understand it as just being intended to provide the student with an opportunity to plan programs of study, select appropriate required and elective
classes, and schedule classes in a way that provides the greatest potential for academic success.

The task is still interesting and of moderate size since when planning a program of study and selecting classes, there are quite a lot of things to consider such as prerequisites, course availability, effective course sequencing, work load, and instructor-student match-ups. Later, when the major problems are understood, improvements and extensions can be attempted, and attempts can be made to tackle the advising problem in a more general framework.

**Resources Needed for Academic Advising**

There are a lot of diverse resources that are required to deal with the problem of academic advising (see Table 1).

First of all, one needs the student profile that includes the courses already attended, the corresponding grades, the “interests of the student” concerning the courses to be attended, and perhaps much other information. The part of the profile consisting of the courses already attended, the corresponding grades, and so on, is maintained by the university administration in appropriate databases to which the access is restricted to some administrators. The part of the profile consisting of the “interests of the student” concerning the courses to be attended exists actually only in the head of the student and should therefore be asked for from the student before advising is performed. However, attempts may be made to let the system learn the interests, for example, by monitoring the student or/and looking at his profile.

The second resource needed for solving the problem of academic advising are the courses that are offered in the semester for which advising is needed. This information is as well maintained by the university administration in appropriate Web sites and is accessible for everyone.

The third resource needed for solving the problem of academic advising is expertise. Expertise is the extensive, task-specific knowledge acquired from training, reading, and experience. It is the knowledge that allows experts to make better and faster decisions than non-experts when solving complex problems. It consists of facts, theories, as well as rules and procedures about a problem area.

For the problem of academic advising this type of expertise may be referred to as the university regulations concerning academic advising. They consist of all the details concerning courses, programs, and curricula such as names/codes of courses, in which term (fall, spring, etc.) courses are offered, prerequisites for courses, descriptions of courses, teaching and assessment methods, statuses (compulsory, elective, etc.) of courses.

---

**Table 1. Information resources for the problem of academic advising**

<table>
<thead>
<tr>
<th>Student Profile</th>
<th>Courses Already Offered</th>
<th>Other Information</th>
<th>Courses Offered</th>
<th>Expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interests of the Student</strong></td>
<td><strong>Courses Offered</strong></td>
<td><strong>Other Information</strong></td>
<td><strong>Documented knowledge:</strong></td>
<td><strong>Knowledge of a human expert:</strong></td>
</tr>
<tr>
<td>Entered by the student</td>
<td>A II information available about courses offered in the semester for which advising is needed</td>
<td>A II information concerning courses already attended by the student</td>
<td>A II information concerning courses already attended by the student</td>
<td>Heuristics, strategies, meta-knowledge, etc.</td>
</tr>
<tr>
<td>Attempts can be made to learn them by monitoring the user and/or looking to his profile</td>
<td>A II information concerning courses already attended by the student</td>
<td>A II information concerning courses already attended by the student</td>
<td>A II information concerning courses already attended by the student</td>
<td>More complex than documented knowledge</td>
</tr>
<tr>
<td>Frequently changed</td>
<td>Stored in a database</td>
<td>A II information concerning courses already attended by the student</td>
<td>A II information concerning courses already attended by the student</td>
<td>Expertise is gained through a learning process</td>
</tr>
<tr>
<td></td>
<td>A II information concerning courses already attended by the student</td>
<td>A II information concerning courses already attended by the student</td>
<td>A II information concerning courses already attended by the student</td>
<td></td>
</tr>
</tbody>
</table>
number of credit hours of courses, definitions and implications of curriculum terminology (e.g., cross-listing, interdisciplinary, etc.), established programs, established tracks, curriculum total credit hours and their distribution (e.g., compulsory vs. elective), titles of academic degrees offered by the university, and so on. This kind of information is provided by the university and is accessible for everyone. It is published in Web pages, booklets, and in many other different forms such as printouts and announcements. The sources of knowledge are many; however, the primary source will be a human expert. A human expert should possess more complex knowledge than can be found in documented sources. This kind of knowledge includes rules (heuristics) of what to do in a given problem situation, global strategies for solving specific problems, and meta-knowledge.

Because, as mentioned earlier, knowledge acquisition is the major “bottleneck” in developing expert systems, we will approach a solution to gaining the expertise needed for academic advising by adopting a machine learning paradigm.

Why a Multi-Agent System?

First of all, academic advising is intended to be a good domain and problem to test the adequacy of the multi-agent paradigm for dealing with information customization. It provides a complex and dynamic environment and constitutes a wonderful experimental test-bed for investigating the issue. Conversely, dealing effectively with the problem of academic advising will require a multi-agent system. The multi-agent paradigm seems to be appropriate and even superior to other approaches such as a traditional distributed database approach enhanced with an intelligent user interface for the reasons elaborated in the following.

The resources needed for academic advising are physically distributed and dynamic:

- **Their content may change:** It is for example frequent that, at the beginning of a semester, some of the offered courses are canceled, and some other ones are added. The profile of the student also changes frequently, for example when the grades of the exams taken are entered. Also the interests of the student concerning the courses to be attended may change. As an example, students usually want to enroll themselves in courses that are attended by some of their friends. Changes to the university regulations, in contrast, especially those concerning the advising process like the curriculums for the different majors, are comparatively less frequent.
- **Their form (structure) may change:** They may be available via a Web page, an intelligent agent, a data base, a legacy system, and so on.
- **Their location may change:** Existing ones may be moved and new ones may be incorporated.

Hence, dedicating a separate intelligent agent to each individual resource for coping with all of its peculiarities will have many advantages such as:

- **Reduction of the scope of changes:** There is no need to change the whole system when changes concern only a specific resource. In this case, only the concerned agent is changed.
- **Easy incorporation of new resources:** Only one new agent is needed for each new resource.
- **Easy extension and improvement of the system:** We may think, for example, of a self adjusting system, that is, the system itself searches for appropriate resources and each time a resource is identified, it is wrapped with an appropriate agent.
DETAILS OF THE MULTI-AGENT-BASED SOLUTION

MASACAD, the multi-agent system described in the following, offers a service to a student who needs academic advising, that is, he wants to know which courses he should enroll himself in for the coming semester. The service is currently available only for computer science students. In the current version, no attempt is made to learn the student profile, that is, the part the profile consisting of the interests of the student regarding the courses to be attended. The interests are therefore asked for from the student before advising him. The focus in the current version is on the individual agents of the system, on how they cooperate to solve the problem, and on how one of them, namely, the “Learning Agent”, learns to perform the advising process by adopting a supervised learning solution using neural networks (see “Advising Procedure”).

As a solution to the problems of network communication, we use Bee-gent (Bonding and Encapsulation Enhancement Agent) (Kawamura, Hasegawa, Ohsuga, & Honiden, 2000), a communication framework based on the multi-agent model. The Bee-gent framework is comprised of two types of agents. “Agent Wrappers” are used to agentify (i.e., providing an agent interface) existing applications, while “Mediation Agents” support inter-application coordination by handling all communications. The mediation agents move from the site of one application to another where they interact with the agent wrappers. The agent wrappers themselves manage the states of the applications they are wrapped around, invoking them when necessary. MASACAD consists of 3 “Agent Wrappers” (Learning Agent, Data Base Agent, WWW Agent) and one “Mediation Agent” (Searcher).

System Architecture of MASACAD

MASACAD consists of a mediation agent that provides the information retrieving service (Searcher) and of a “User System”, a “Grading System”, and a “Course Announcement System” that are technically applications, each of them existing within an agent wrapper, that is, they represent different agents of the advising system (see Figure 1).

The student (user) uses a Graphical User Interface (GUI) to utilize the service of the advising system. When the student enters his identifier and

Figure 1. Process flow through MASACAD

GUI: Graphical User Interface
JDBC: Java DataBase Connectivity
IRS: Information Retrieval Service
DB: Data Base
URL: Uniform Resource Locator
password, the agent wrapper for the “User System” application creates a mediation agent (Searcher), which migrates to the agent wrappers of the “Grading System” application and of the “Course Announcement System” application to retrieve the needed information (“courses successfully taken by the student with the corresponding grades” and “courses offered in the semester for which advising is needed”). After that, the student has to enter his interests by choosing courses among those that are offered in the semester for which advising is needed by checking some boxes showing the course names. When the student finishes entering his interests, he has to just click on the “Get Advice” button to obtain a list of courses sorted in descending order of importance in which he is advised to enroll himself in.

In the following, each of the individual components of the system is be described in detail.

### Grading System

The application “Grading System” is a data base application for answering queries about the students and the courses they have already taken. The agent wrapper for the application “Grading System” is responsible for invoking the database by using JDBC (Java DataBase Connectivity), transforming requests for information about the students and courses into queries to the database system, collecting the results of the queries, and finally replying these results to the mediation agent. Table 2 summarizes the steps performed by the agent wrapper of the application “Grading System”.

### Course Announcement System

The application “Course Announcement System” is a Web application for answering queries about the courses that are expected to be offered in the semester for which advising is needed. The agent wrapper for the application “Course Announcement System” takes on the invocation of the Web application. It is responsible for initiating communication with the Web application, transforming requests for information about the courses into queries to the Web application, collecting the results of the queries, and finally replying these results to the mediation agent. Table 3 summarizes the steps performed by the agent wrapper of the application “Course Announcement System”.

### Mediation Agent

The mediation agent realizes services by interacting with the agent wrappers on the basis of conversations (sending and receiving XML [eXtensible Markup Language] messages). When the mediation agent migrates, it carries its own program, data and current state. Frequency of communication is reduced compared to a purely message-based system and network loads are decreased largely because communication links can be disconnected after the launch of the mediation agent. Processing efficiency is improved because the mediation agent communicates with the applications locally. The behavior of the mediation agent is described in Table 4.

### User System

The agent wrapper for the “User System” application creates a mediation agent (Searcher), which migrates to the agent wrappers of the applications “Grading System” and “Course Announcement System” to retrieve the needed information. After that, the “GUI Application” is contacted to give the opportunity to the student to express his interests. The “Advising Procedure” (see next
The agent wrapper for the “User System” application also monitors changes in the offered courses and in the student profile and alerts the user by sending an e-mail. Table 5 summarizes the steps performed by the agent wrapper of the application “User System”.

### The Advising Procedure

The “Advising Procedure” is focused on advising computer science students. Computer science students, in order to complete their degree, must accomplish during their study a total of 132 credit hours by choosing at each term courses among the 85 courses that are specified by the curriculum. Some of these courses are compulsory, others are elective. The choice should of course occur according to the university regulations and in a way that provides the greatest potential for academic success as seen by a human academic advisor.

Taking into account the adequacy of the machine learning approach for gaining human expertise added to the availability of experience with advising students makes the adoption of a paradigm of supervised learning from examples obvious. Back-propagation is the best-known example of a supervised neural network training algorithm Patterson (1996) and is therefore used to implement the “Advising Procedure”. The known information (input variables) consists of the profile of the student and of the offered courses grouped in a suitable way. The unknown information (output variables) consists of the advice expected by the student. In order for the network to be able to infer the unknown information, prior training is needed. Training will integrate the expertise in academic advising into the network.

### Settings of the Neural Network Solution

For the problem of academic advising, information about 250 Computer Science students in different stages of study was collected, which means that 250 examples were available for the learning procedures. The 250 examples were used to form three sets: a training set containing 150 examples, a selection set consisting of 50 examples, and a test set consisting of 50 examples. In an attempt to reduce the number of parameters that should be tuned, we decided to use two hidden layers.

To implement the network, a slightly modified version of the package developed by Mark Watson (2005) was used.

Each one of the 250 examples consists of a pair of input-output vectors. The input vector summarizes all the information needed for advising a particular student. It consists of 85 components.
Table 4. Mediation agent

- Receive request for information retrieval from agent wrapper for “User System”
- Move to agent wrapper for “Grading System”
- Judge the migration result:
  - If failure: end
  - If success:
    - Request for retrieval of (course, grade) pairs for the given student ID
    - Receive the reply
    - Move to agent wrapper for “Course Announcement System”
    - Judge the migration result:
      - If failure: end
      - If success:
        - Request for retrieval of IDs of offered courses for the given semester, college, and subject
        - Receive the reply
        - Move to agent wrapper for “User System”
        - Judge the migration result:
          - If failure: end
          - If success:
            - Report retrieved results to agent wrapper for “User System”
            - End

Table 5. Agent wrapper for “User System”

- When GUI event occurs (student enters ID, Password, and Email address the first time, or asks again for advice by clicking the button “Get Advice Again”) or when periodic monitoring of changes in student profile and offered courses is due:
  - Create mediation agent
  - Request mediation agent to perform information retrieval
  - Receive results from mediation agent
  - If information retrieval was initiated by a GUI event then:
    - Output information concerning the offered courses to the GUI (the “GUI Application” represents this information appropriately (boxes that can be checked by the student))
    - When GUI event occurs (student clicks on “Get Advice” button):
      - Prepare input (input vector stored in a text file) for “Advising Procedure”
      - Invoke “Advising Procedure”
      - Process results (the results are found in a text file) and output them to GUI
    - Else (periodic monitoring): compare retrieved information with old version and alert the user via Email in the case that something has changed

Each of them containing a number for one of the 85 courses in the curriculum. The number consists of the grade earned if the student has already passed the given course. Otherwise, the number encodes the availability of the course in the semester for which advising is needed and the interest of the student in the course. The output vector encodes the final decision concerning the courses in which the student actually enrolled himself in based on the advice of the academic advisor. This happens by assigning priority values to each of the 85 courses in the curriculum.
Learning Phase

The aim of the learning phase was to determine the most suitable values for the learning rate, the size of the network, and the number of training cycles that are needed for the convergence of the network.

**Learning Rate:** To determine the most suitable learning rate, experiments with 10 network configurations 85-X-X-85 with X in \{3, 5, 7, 9, 20, 30, 50, 85, 100, 120\} were performed (85-X-X-85 represents a network with 85 input neurons, X neurons in each of the two hidden layers, and 85 output neurons). For each of these networks, experiments with the five learning rates: 0.01, 0.25, 0.50, 0.75, and 1.0 were conducted. In each of these 50 experiments, the network was allowed to learn for a period of 1,000,000 cycles. After each epoch of 50,000 cycles the average selection error for the 50 pairs from the selection set was calculated. From the 50 experiments it was clear that the learning rate 0.01 is the most suitable one for this application because it produces the smallest selection error in most of the cases and more importantly, it causes the selection error to decrease continuously which forebodes the convergence of the network.

**Network Size:** Figure 2 shows the average selection error for the 50 pairs from the selection set plotted as a function of the number of training cycles for the 10 different network configurations 85-X-X-85 with X in \{3, 5, 7, 9, 20, 30, 50, 85, 100, 120\}. In all 10 cases the network was allowed to learn for a period of 1,000,000 cycles and the learning rate was set to 0.01. The network configurations 85-X-X-85 with X in \{50, 85, 100, 120\} seem to work best in combination with the learning rate 0.01. To determine which one of them is more suitable, a longer learning period is needed. The results for a learning period of 10,000,000 cycles are illustrated in Figure 3. The configurations 85-100-100-85 and 85-120-120-85 cause the selection error to decrease continuously and reach a very satisfactory level. Both networks have similar performances. However, following the standard scientific precept that, all else being equal, a simple model is always preferable to a complex one, we can also select a smaller network in preference to a larger one with a negligible improvement in selection error. The network configuration 85-100-100-85 is therefore retained as the best network configuration (in terms of selection error).

**Number of Training Cycles:** After a given number of training cycles, when the selection error and consequently the performance reaches a satisfactory level and when the improvement becomes negligible, learning can be stopped and the network producing the smallest selection error can be retained as a solution for the problem of academic advising. In our case, this happens with the learning rate 0.01 and the network configuration 85-100-100-85 after a training period of 800,000 cycles.
Testing Phase

The final model was tested with the test set data. For 42 of the 50 test cases (84 percent) the network’s actual output was exactly the same as the target output, that is, the network suggested the same courses in the same order as specified by the test examples. In the remaining eight test cases, the target courses were always present at the beginning of the course list produced by the network. However, the network proposed some additional courses. The courses proposed additionally occur always at the end of the course list. This, in addition to the fact that the system is an advisory one, makes these errors tolerable. In four of the eight cases the additional courses were correct choices. In the other four cases some of the additional courses were wrong choices because the student has not yet taken their prerequisite courses.

Benefits and Limitations of the Neural Network Solution

The research reported showed the applicability of the neural network approach to academic advising. The rewards of a successful “marriage” of neural networks and expert systems are too enticing. The back-propagation network performed favorably and seems interesting and viable enough to be used to automate the process of knowledge acquisition, usually looked upon as “the bottleneck problem” of expert systems applications, and help in extracting human expertise, needed for solving the problem of academic advising. In this vein, a possible way to incorporate neural networks into expert systems was described. This would smooth the way for a real and significant “fusion” of neural networks and expert systems. With a network topology of 85-100-100-85 and systematically selected network parameters (learning rate of 0.01, 8000000 training cycles), the multiple-layered, fully connected back-propagation network was able to deliver a considerable performance.

Despite the promising results demonstrated in various applications, back-propagation network development still requires extensive experimentation, parameters selection, and human judgment. Developing a good neural network is not a trivial task as the technique currently stands. Some interesting research questions about adopting neural networks remain to be answered. For example, can we develop a systematic and simple methodology and blueprint for neural network modeling and analysis? Can we automate the selection of topology and parameters? Can we let the network “learn” incrementally without major network re-training when new information is discovered (e.g., curriculum changes)? Incremental learning would require the exploration of other neural network models, such as fuzzy ARTMAP (Carpenter, Grossberg, Markuzon, Reynolds, & Rosen, 1992) and ARAM (Tan, 1995).
SYSTEM EVALUATION

To evaluate the advising multi-agent system in real academic environment, 20 computer science students in different stages of study were involved. Each of them was asked to use the system to get academic advice for the coming term. Most of the problems that occurred at the beginning were of technical nature, mainly concerning the communication over the network. It took a while to deal with this kind of problems and to obtain a relatively stable system.

The advice delivered by the system in each of the 20 cases was analyzed carefully by the concerned student together with an academic adviser with the aim of evaluating its suitability and detecting possible errors. The results were very encouraging. In none of the 20 cases was the advice found to contain errors, that is, to contain courses that, if chosen, will violate the university regulations such as those concerning course prerequisites. Also, all of the results (suggested courses) were judged by the concerned students and academic advisers to be suitable and in most of the cases even very appropriate to be taken by the students. Actually, 15 of the 20 students took during the term for which advising was needed exactly the courses suggested by the advising system. In one of the remaining five cases, the student, because of her poor health, was later (at the beginning of the term) forced to drop three courses and to take only two of the five suggested ones. In another case, the student didn’t take any course during the term because of leave of absence for the whole term. In the other three cases, each of the students exchanged a suggested course (the same course in all three cases) with another one. Later, it became clear that the three students were in a clique and that one of them didn’t care for the instructor of the course, so they moved all to the other course.

BENEFITS AND LIMITATIONS OF MASACAD

MASACAD, in addition to providing a solution for the problem of academic advising, was also used to demonstrate, on an example, the suitability of the multi-agent paradigm for dealing with information customization and personalization. Information customization and personalization provides users with relevant content based on their needs and interests and this is exactly what MASACAD does. It provides the user (student) with an interactive, personal Web experience and customizes the information to the user’s desire and needs. The preferences (profile) of the user determine the behavior of the system. Using the multi-agent approach combined with other techniques, we were able to devise an elegant solution to a problem that depends on distributed information existing in differing formats. Conversely, as mentioned earlier, the distributed and dynamic nature of today’s information resources makes the adoption of the multi-agent approach necessary.

However, there are some other important characteristics of information customization systems that were not demonstrated on this example. In general, in an information customization system there should be some strategy for discovering the location of potential useful information for example among the whole Web, among a reasonable part of the Web, among a set of databases, or among any other heterogeneous set of resources. Also it should be possible to automate the creation of the mediation agent and the agent wrappers, that is, the creation of the whole multi-agent system. The multi-agent system to be created depends on the resources to be invoked and on their locations. The automation task may therefore be very difficult: the resources can be numerous and may have very different formats (they may be unstructured Web pages, databases, agents) and
this complicates the creation of the agent wrappers (how to wrap an unknown application?), as well as the creation of the mediation agent (which route to take to navigate to the different agent wrappers, how to communicate with the different agent wrappers).

FUTURE DIRECTIONS OF WORK

As mentioned earlier, the MASACAD system in its current version is available only for computer science students. The next step for this work is therefore to focus on extending the system to cope with all other degrees in the university. This will imply that for each of these degrees, data about advising students should be collected in order to construct the training, selection, and testing sets. Also, for each degree, a neural network should be developed which requires extensive experimentation, parameters selection, and performance judgment.

However, before extending the system to cope with other degrees, the system should be tested on a larger scale to detect possible errors, be convinced of its usefulness, and better assess its power. This will be done by trying to involve all Computer Science students (more than 300) in evaluating the system. Before allowing a student to register to any course, the student should get advice from the system and discuss its appropriateness with his human academic adviser.

Another improvement, necessary to think about, is how can we let the neural networks “learn” incrementally without major network retraining when new information that is suspected to affect the process of academic advising is discovered (e.g., curriculum changes)? This will make the system more flexible and avoid the repeated new development of neural networks each time a change occurs.

Another problem that should be considered is how can the system be made flexible against changes in the location of course announcements and student profiles (where do they reside?) and against changes in the type of application that enables access to this information (is it a Web site, a database, an intelligent agent, etc.?). To solve this problem, one can think of automating the multi-agent system, that is, creating the mediation agent and the agent wrappers automatically. This will, of course, require knowing which sites should be wrapped by agent wrappers (i.e., knowing the location of resources), and how this wrapping should be done (based on the type of the site/application).

RELATED RESEARCH

In recent years, the administration of academic advising has undergone radical transformation as technological developments have altered the processes by which information is collected, stored, and accessed; the systems by which communication is enabled; and the structures by which transactions are conducted. Technological innovations have created an abundance of opportunities for new practices and enhanced services “frequently characterized as ‘real-time’, ‘student-centered’, and ‘any time, any place’” (Moneta, 1997, p. 7). The technological environments on many campuses are evolving rapidly and comprise numerous elements: information dissemination, transactional interaction, communications applications, and educational technologies. Steele and McDonald (n.d.) contains an annotated bibliography compiled by George Steele and Melinda McDonald for research related technology and advising.

The problem of course assignment has been studied in the literature from different angles. Advising tools are usually intended to complement the student-advisor relationship. The course-planning consultant (CPC), for example, is a computerized Web-based prerequisite checker developed for students in the General Engineering undergraduate curriculum at the University of Illinois at Urbana-Champaign. It was reported that
the program has been field-proven to have helped prevent costly mistakes when planning future courses (Course Planning Consultant, n.d.).

There are also few learning environments where similar problems, such as learning objects assignment, were addressed using the multi-agent system approach. In Sheremetov and Arenas (2002), for example, personalized learning trajectories are constructed by agents in the learning space based on users’ profiles.

It is surely difficult to compare the performance of MASACAD with these systems because they do not perform exactly the same task. However, the technical novelty of MASACAD seems to lay in its ability to perform academic advising while taking into account not only visible static information but also hidden expertise. MASACAD also demonstrates the capability of exploiting the digital infrastructure, enabled by the online mode of teaching and learning, to extract and infer useful knowledge. This is done in an elegant way through a multi-agent system inspired from human teams and organizations involved in the problem-solving activities.

**CONCLUSION**

Academic advising is an important area in any university, which seldom gets adequate resources. Hence, any system that can support this operation will be worthwhile. The idea of realizing such a system is very interesting and highly related to the current research and development trends. In this paper, MASACAD, a well defined architecture for a multi-agent system for addressing the problem of academic advising was proposed. It uses intelligent agents and neural networks for learning and recommendation. The prototype implementation and preliminary testing show that the multi-agent paradigm, combined with ideas from machine learning, user modeling, and Web mining, would be a good approach for the problem of information customization. It was very interesting to see an e-learning system that is able to learn and react according to the characteristics of the client (in this case the student). The results are encouraging and future work will be focused on improving the system and studying how such simple examples, built with insight, should lead to the identification of key difficulties, useful abstractions, and a general method for solving the problem and revelation of the issues.

**REFERENCES**


Moneta, L. (1997). The integration of technology with the management of student services. *New Directions for Student Services: Using Technology to Promote Student Learning, 78*, 5-16.


*This work was previously published in International Journal of Intelligent Information Technologies, Vol. 2, Issue 1, edited by V. Sugumaran, pp. 1-20, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 3.9
Human Factors Problems of Wearable Computers

Chris Baber
The University of Birmingham, UK

James Knight
The University of Birmingham, UK

ABSTRACT

In this chapter wearable computers are considered from the perspective of human factors. The basic argument is that wearable computers can be considered as a form of prosthesis. In broad terms, a prosthesis could be considered in terms of replacement (i.e., for damaged limbs or organs), correction (i.e., correction to ‘normal’ vision or hearing with glasses or hearing aids), or enhancement of some capability. Wearable computers offer the potential to enhance cognitive performance and as such could act as cognitive prosthesis, rather than as a physical prosthesis. However, wearable computers research is still very much at the stage of determining how the device is to be added to the body and what capability we are enhancing.

INTRODUCTION

There is a wide range of technologies that have been developed to be fitted to the person. Depending on one’s definition of “technology,” this could range from clothing and textiles, through to spectacles, to cochlear implants. The use of these different technologies can be basically summarized as the supplementation or augmentation of human capability, for example the ability to regulate core temperature (clothing), to see (spectacles) or to hear (cochlear implant). One reason why such supplementation might be required is that the current capability does not fit with environmental demands, either because the environment exceeds the limits over which the human body can function or because the capability is impaired or limited. From this perspective, a question for wearable computers should be what are the cur-
rent human capabilities that are exceeded by the environment and require supplementation by wearable computers? In the majority of cases for wearable computers, the answer to this question hinges on communicative, perceptual, or cognitive ability. As Clark (2006) notes, “...the use, reach, and transformative powers of these cognitive technologies is escalating” (p. 2). But the essential point to note is that “Cognitive technologies are best understood as deep and integral parts of the problem-solving systems that constitute human intelligence” (p. 2). Thus, such technologies could represent a form of ‘cognitive prosthesis’ in that they are intended to support cognitive activities, for example, having a camera (performing face-recognition) to advise the wearer on the name of the person in front of them. The immediate challenge is not necessarily one of technology but of cognition. If the technology is ‘doing the recognition,’ the question is raised what is the human left to do? To draw on a commonly cited analogy, spectacles serve as a perceptual prosthesis, that is, to improve or correct a person’s vision. For wearable computers (and the related field of augmented reality), the ‘improvement’ could be to reveal to the person objects that are not present by overlaying an artificial display onto the world. The display could simply take the form of labels or directional arrows, or could be more a sophisticated presentation of moving (virtual) objects. In both cases, the ‘augmentation’ could either enhance the person’s understanding of the environment or could substitute this understanding.

In terms of communication, mobile telephones and MP3 devices contain significant computing power such that they can easily be considered as being computers, albeit with limited functionality. These devices can be worn, for example MP3 players can be worn on the upper arm, attached to belts or neck-straps, or placed in hats, and mobile telephones can be attached to belts. Furthermore, with the wireless (Bluetooth) headset, the user interface of a mobile telephone can be worn on the ear at all times. Thus, both can be always present and both can be considered part of the person. A definition of wearable computers ought to, at least, allow differentiation from devices that can slip into the user’s pockets (if this technology is to be treated as a new area of research and development). Two early definitions of wearable computers, from Bass (1996) and Mann (1997), emphasize that wearable computers are designed to exist within the corporeal envelope of the user and that this makes them part of what the user considers himself or herself. In many respects this allows an analogy to be drawn between wearable computers and prosthetic devices. Having something added to the body, whether externally, such as spectacles, artificial limbs, hearing aids, or internally, such as pace-makers, or cochlear implants, changes the performance of the person and (for external prosthesis) the appearance of the body. This now becomes a very different concept from the mobile telephone, MP3 player and the computer that we traditionally encounter. This raises all manner of interesting questions relating to physical, perceptual, and cognitive aspects of human factors, as well as a whole host of emotional aspects of wearing devices (for the wearer and the people with whom they interact). At the moment, there remains a gap between what a wearable computer is intended to be and what mobile telephones and MP3 players currently are. This gap can best be considered as a form of perceptual and cognitive prosthesis, in which the wearer’s ability to view the world, retrieve pertinent information, and respond to environmental demands are enhanced by the technology across all aspects of everyday life. At present mobile telephones and MP3 players are able to be tailored (by the user) and can deal with a limited set of situations (relating to communications or music playing) but do not fill the specification that one might have for a wearable computer. The basic difference between a wearable computer and these other technologies lies in the question of how well the user can interact with both the device and the environment.
simultaneously. Obviously, listening to an MP3 player or speaking on a mobile telephone can be performed while walking through an environment. However, performing control actions on the devices can be sufficiently demanding to draw the user’s attention from the environment. The ideal wearable computer would allow the user to manage attention to both device and environment. In other words, it should allow the user to manage both foreground and background interaction (Hinckley et al., 2005).

**FORM-FACTOR AND PHYSICAL ATTACHMENT**

Moore’s law continues to guarantee that the processors will get smaller, and work in the field of micro-electrical-mechanical systems (MEMS) shows how it is possible to create application specific processors that are small enough to be incorporated into buttons on clothing or into jewelry (Figure 1a).

Furthermore, it is feasible to assume the widespread development of general purpose processors, such as the mote (and similar) concept, that combine low power with sufficient processing capability to deal with a number of different sensors (figure 1b).

One direction for wearable computers is that the miniaturization of technology will mean it is possible to implant processors under the skin (or embed them into clothing). This is not a particularly novel idea as the medical world has been experimenting with mechanical implants for many years and has, over the past decade or so, developed digital implants that can, for example, regulate drug administration or improve hearing ability. While the development of such technology is exciting and likely to lead to fascinating discoveries, it lies somewhat outside the remit of this chapter. The problem is how can these extremely small devices support interaction with the person wearing (or supporting) them? To a great extent, the human simply has these devices fitted into them. While this represents the logical extension of the argument that the wearable computer exists within the wearer’s ‘corporeal envelope,’ it does preclude the possibility of developing means of interacting with the device. Indeed, as technology continues to get smaller, ever more difficult the challenge of supporting human-computer interaction becomes, for example, how can one read displays with very small font or press buttons that are much smaller than the human finger? This could mean that we should redefine our concept of interaction, for example, one could argue that ‘interaction’ should be any activity that the person performs, that is sensed by the device and that allows the device to make a response. A potential problem with such a concept is that the person might not be able to selectively control what the device senses or to fully understand why the device is behaving in the manner that it is. This

*Figure 1a. MEMS device*  
*Figure 1b. Intel’s Mote prototype*
means that one either locks the human out of the interaction (in which case one has a sensor system that acts on the environment and which coincidentally affects people in that environment) or one must develop different ways in which the person can understand and manage the behavior of the device. This is not a trivial problem and one can point to many developments in the ubiquitous and pervasive computing domain in which the role of the human in device performance is merely one of passive recipient of device activity. From the point of view of human-computer interaction, this will not only lead to the ‘irony of automation’ but also to frustration, annoyance and a perceived lack of control. In other words, who wants to live in a world that is basically run by the descendants of the Microsoft Office Paperclip?

**Physical Effects of Wearing Computers**

We might expect a computer to consist of a processor and storage, some form of power, a display, and an interaction device. For wearable computers, each of these components can be separated and worn on different parts of the body. Thus, as the prototypes in Figure 2 illustrate, the processor and storage could be incorporated into a single unit and mounted on the waist or the back, the display could be mounted on the head (either near the eyes or the ears, depending on the feedback provided to the wearer), the power could be located near the processor unit, and the interaction device could be mounted within easy reach of the hand (if manual control) or the mouth (if speech input). These prototypes are typical of much contemporary work on wearable computers in terms of the relative size and placement of components.

Wearable computers represent a load on the person and consequently can affect the physical activity of the person (Zingale et al., 2005). Legg (1985) proposes that the human body can carry loads at the following points: head, shoulder, back, chest, trunk, upper arm, forearm, hands, thighs, feet, a combination of these, and aided (i.e., by pulling, pushing or sharing a load). A glance through the literature of wearable computers shows that most of these sites have been experimented with in various designs. Gemperle et al. (1998) offer the term ‘wearability’ to describe the use of the human body to physically support a given product and, by extension, the term ‘dynamic wearability’ to address the device being worn while the body is in motion. Given this notion of wearability, there is the question

*Figure 2. Wearing prototype wearable computers*
of where a device might be positioned on the body, and, once positioned, how it might affect the wearer in terms of balance, posture and musculoskeletal loading. Given the notion of dynamic wearability, there are the questions of how the device will be carried or worn, how this might affect movement, and how this might lead to either perceptions of differences in movement patterns, physiological strain or psychological stress on the wearer. These changes can be assessed, using subjective self-report techniques (Bodine & Gemperle, 2003; Knight et al., 2002, 2006; Knight & Baber, 2005) and through objective analysis (Nigg & Herzog, 1994).

**Energy Expenditure, Muscle Activity and Using Wearable Computers**

An attached load to the body will have a direct affect on the energy expended by the body, as the muscles burn more energy to generate force to counteract the weight of the load. This situation sees a wearer of a wearable computer potentially increasing their overall energy expenditure to overcome inertial changes. Thus, any additional weight on the body can create demands on the musculoskeletal system to support and move the object. In addition, the kinematics of posture and movement can compound the loading effects of a wearable computer. Knight and Baber (2004) report the following about head-mounted loads, such as head mounted displays: (1) different head postures have a measurable effect on musculoskeletal loading; (2) the heavier the frontal load (as is representative of the loading of a HMD), the greater the muscle activity required to keep the head in a fixed position and; (3) in neutral or extended postures, the wearer can sustain larger loads than if the head is in a flexed or rotated head position. Thus, it is not simply a matter of reducing load of head-mounted equipment, but one must determine the posture that the wearer of such equipment is likely to adopt. An alternative location for display technology is on the forearm. In their study of pointing devices, Thomas et al. (1998) demonstrate that for general purpose activity, the forearm would be an appropriate place to mount the device. This location is attractive as it allows the wearer to easily move the display into the field of vision. However, Knight and Baber (2007) have questioned this location or at least raised concerns. Recording shoulder and upper arm muscle activity and measuring perceptions of exertion while participants interacted with arm mounted computers of different weights they found that the mere act of holding the arm in an appropriate posture to interact with an arm mounted computer was sufficient to exceed recommended levels of muscle activity for sustained activity. In addition it induced symptoms of fatigue after only two minutes where the addition of weight in the form of the mounted technology compounded this physical effect.

**Reducing Size and Separating Components**

The discussion so far has suggested ways in which weight and placement of loads on the body can lead to problems. The implication is that most wearables take the form factor of the ‘brick-on-the-back’ (as illustrated by Figure 2). It is possible to reduce the size of components further by removing the interaction and display components. Thus, the Bluetooth headset for a mobile phone effectively reduces the user interface (for talking on the phone) to a unit that clips on the ear. Various MP3 players can be worn on the upper arm while jogging (see comment about usability). Alternatively, there has been much interest in the use of ‘active badges’ that signal a person’s location to a network (Want et al., 1992). These devices can be made small enough to be worn as badges or incorporated into items of clothing, for example Schmidt et al. (1999) report a device, mounted in the wearer’s tie, which detects changes in ambient sound levels or wearer movement. In these examples, the ‘on-body’ part of the computer system is basically a sensor and/or transmitter and/or receiver that link the wearer
to a network of other technologies. BodyMedia have developed a device that places sensors on the person’s upper arm in order to record data relating to everyday activity (in much the same way that an MP3 player is a data display device) and needs to be connected to a computer in order to process the tunes on it (see Figure 3). However, it makes sense to ask whether reducing the size of the devices will reduce or even eliminate these problems. Physical discomfort may arise from pressure of the device on the body. The pinnar is very sensitive and many users complain of discomfort around the ear from using ear plugs and ill-fitting head phones. HMDs place pressure around the forehead and crown. Even the small displays produced by Microoptical (see Figure 4) can cause discomfort around the ears and nose; especially given that the glasses they come with are hard plastic, not specifically fitted for the individual wearer. Items attached around the arm, specifically during physical activity (i.e., for jogging), have to be attached tightly so that they do not bounce against the arm or slip down, as such they may result in discomfort and result in vasoconstriction leading to sensations of numbness and tingling in the lower arm and hand, not to mention that the display is not easily viewable when the device is in position which raises usability issues.

**Perceptual Impacts of Wearing Computers**

A wearable computer can be considered a ‘perceptual prosthesis’ in the ways that it can provide additional information to, or enhance perceptual capability of, the wearer. Information provision can simply mean allowing the wearer access to information that is not currently present in the world, for example through a visual or audio display. The underlying concept in many of these applications is that the wearer’s perception of the world can be augmented or enhanced (Feiner et al., 1998). Before considering this argument, we can consider the MP3 player as a device that can present media that is added to the world (in the form of a personal soundtrack to the person’s activity). To some extent this could be considered as form of augmenting perception. What makes augmented reality different from simply displaying information is that (in most systems) the information is presented in accordance with the context in which the person is behaving. Such information could be in the form of an opaque visual display or could be visual information overlaid onto the world. In basic systems, the information can be called up by the wearer, or pushed from another source, for example a radio or telephone link. In more sophisticated systems, the information is presented on the basis of the computer’s interpretation of ‘context.’

While augmented reality displays could be beneficial, there are potential problems associated with the merging of one source of information (the computer display) with another (the world). Contemporary wearable computers tend to combine monocular head-mounted displays (see Figure 4) with some form of interaction device, for example keyboard, pointing device or speech. NRC (1997) point out that monocular head-mounted displays could suffer from problems of binocular rivalry,
that is, information presented to one eye competes for attention with information presented to the other, which results in one information source becoming dominant and for vision to be directed to that source. A consequence of this phenomenon is that the wearer of a monocular display might find it difficult to share attention between information presented to the ‘display eye’ and information seen through the ‘free eye.’ This problem is compounded by the field of view of such displays.

The field of view for the commercial monocular HMDs ranges from 16°-60°, which is considerably less than that of normal vision (around 170° for each eye, Marieb 1992). Narrow field of view can degrade performance on spatial tasks such as navigation, object manipulation, spatial awareness, and visual search tasks. Restrictions on field of view will tend to disrupt eye-head coordination and to affect perception of size and space (Alfano & Michel, 1990). One implication of a restricted field of view is that the wearer of a see-through HMD will need to engage in a significant amount of head movement in order to scan the environment (McKnight & McKnight, 1993). Seagull and Gopher (1997) showed longer time-on-task in a flight simulator when using a head down visual display unit than with a head-mounted, monocular display. Thus, it appears that a monocular display might impair performance. Apache helicopter pilots currently wear monocular, head-mounted displays and a review of 37 accidents concluded that 28 of these accidents could be attributed to wearing of the display (Rash et al., 1990).

**Computer Response to Physical Activity**

While it is possible that wearing technology affects user performance, it is also possible that the physical activity of the person can affect the computer, for example through the use of sensors to recognize actions and use this recognition to respond appropriately.

Given the range of sensors that can be attached to wearable computers, there has been much interest in using data from these sensors to define and recognize human activity. This has included work using accelerometers (Amft et al., 2005; Junker et al., 2004; Knight et al., 2007; Ling & Intille, 2004, Van Laerhoven & Gellersen, 2004; Westeyn et al., 2003) or tracking of the hand (Ogris et al., 2005). The approach is to collect data on defined movements to train the recognition systems, and then use these models to interpret user activity. At a much simpler level, it is possible to define thresholds to indicate particular postures, such as sitting or standing, and then to use the postures to manage information delivery (Bristow et al., 2004). The use of data from sensors to recognize human activity represents the merging of the research domains of wearable computers with that of pervasive computing, and implies the recognition not only of the actions a person is performing but also the objects with which they are interacting (Philipose, et al., 2004; Schwirtz & Baber, 2006).

An area of current interest for the wearable or ubiquitous computing communities is the interpretation of human movement related to maintenance and related activity. The artifacts with which the person is interacting could be instrumented. For example, a simple approach is to fit switches on
components (in a self-assembly furniture pack) and for the user to depress the switches when each component is handled (Antifakos et al., 2002). A less intrusive approach would be to fit radio frequency identification person is using (Schwirtz et al., 2006) and to use the activation of these tags to infer user activity. This requires that either the components or the tools be adapted to the task. The decreasing costs of RfID suggest that, within a few years, tags will be universally used in a wide range of consumer products. Alternatively, the activity of the user could be taken to develop predictive models in order to infer the activity that is being performed. For some activities it might be sufficient to use data from very generic sensors, such as microphones, to collect data to define actions (Ward et al., 2006), while in other it might be necessary to rely on more specific sensors, such as the use of accelerometers to define movements that are characteristic of assembly and maintenance tasks (Schwirtz & Baber, 2006; Westyn et al., 2003). Steifmeier et al. (2006) show how tracking the motion of the hand (e.g., using ultrasonic tracking and inertial sensors) can be used to define specific types of movement that relate to maintenance tasks, such as spinning a wheel or rotating the pedals, unscrewing a cap or using a bicycle pump. In a related domain, there is some interesting work on the collection of activity data relating to nursing, for example through a combine infra-red proximity sensing and accelerometers (Noma et al., 2004).

**USING WEARABLE COMPUTERS**

Given that a wearable computer provides a means of ‘anytime, anywhere’ access to information, most forms of cognitive activity have been mooted as possible areas that can be supported. In this section, we consider three areas: (i.) supporting memory; (ii.) supporting navigation; and (iii.) information search and retrieval.

**Supporting Memory**

A common example used to illustrate the benefits of a wearable computer is what can be termed the ‘context-aware memory.’ Imagine you are attending a conference (or any large social gathering) and having to remember someone’s name. Systems using some form of badge or face-recognition have been proposed to help with such situations; the computer would register the person and provide you with name and some additional details about the person, for example when you last met, what research interests are listed on their web-page, where they work, and so forth. There has been little research on whether and how these systems improve memory, and this example points to a possible confusion between supporting processes involved in recalling information from memory and the provision of contextually-relevant information. An associated question is whether wearable computers (particularly having a head-mounted display) positioned on the eye can have an impact on recall. The analogy is with the tourist watching a parade through the lens of a video camera—does the act of recording something weaken the ability to process and recall information? Baber et al. (2001) use a search task coupled with surprise recall to show that, in comparison with not using any technology, participants using a digital camera and wearable computer conditions showed lower performance, and that overall the wearable computer showed the biggest impairment in recall. There are many reasons why interruption at initial encoding can limit the ability to remember something, and the question is whether the head-mounted display serves to interrupt encoding; either due to distraction (with a host of information appearing on the screen), or through limitations of field of view, or for some other reasons.
Navigation and Way-Finding

Wayfinding requires people to travel through the world in order to reach specific locations. Thus, there is a need to manage both the act of traveling (for wearable computers this usually consists of walking) and relating a view of the world to the defined location. Sampson (1993) investigated the use of monocular, head-mounted displays for use when walking. Participants were presented with either ‘spatial’ or alphanumeric information and required to traverse paths with or without obstacles. In general, participants performed equally well when standing or when traversing paths without obstacle, but were significantly worse when obstacles were present. Thus, the need to maintain visual attention on display and environment can be seen to impair performance which suggests that the amount and type of information which can usefully be presented on such displays needs to be very limited. As global positioning systems (GPS) become smaller, cheaper, and more accurate, there has been an increase in their application to wearable computers. Seager and Stanton (2004) found faster performance with a paper map than GPS, that is routes completed faster. This was due, in part, with the increase in time spent looking at the GPS display and the number of updates made on the digital view (presumably because the GPS was moving a marker along the map to show the participants location). Participants also were less likely to orientate the digital view in the direction of travel (possibly because they might have assumed that the display would orient to their direction of travel rather than North up). Studies into the effect of perspective views, that is aligning a map with direction of travel, have not shown significant performance advantage over 2D views to date (Suomela et al., 2003), although this approach does seem beneficial in systems that support wayfinding in moving vehicles (Aretz, 1991). Systems that overlay routes onto the head-mounted display (Figure 5) could also assist in simple wayfinding tasks.

The use of visual support for navigation requires the user to divide attention between the environment and a visual display. In terms of the background or foreground of activity (discussed by Hinckley et al., 2005), this essentially places all the tasks in the users foreground. It is possible that navigation prompts could be provided using auditory cues. In simple terms, the heading could be indicated by varying the parameters of simple ‘beeps,’ for example changing in pitch or in intensity as the person deviates from a path (rather like auditory glide-slope indicators in aircraft). More recent developments have replaced the simple ‘beeps’ with music. In this way, a more subtle

Figure 5. Displaying a route overlay on a head-mounted display
(background) form of cueing can be achieved to useful effect. The music could be manipulated to vary quality with deviation from a path, for example through distortion (Strachan et al., 2007, 2005), or through modifying the panning of music in stereo presentation (Warren et al., 2005).

Finding and Retrieving Information

Having the ability to access information as you need it is a core concept of wearable computer research. Early examples had the user enter queries and view information on a monocular display. This approach of having the user ask for information has been superseded by having the computer push information, on the basis of its interpretation of context. In such examples, the benefit of wearing the computer comes from its permanent presence and state of readiness, allowing access either of data stored on the computer or via the World Wide Web. Rhodes and Starner (1997) describe the ‘re-membrance agent,’ which monitors the information that a user types into a computer and makes associations between this information and data it has stored. Obviously this creates an overhead on the user, in terms of the need to type information into the computer. However, it is only a relatively small modification to replace the typed entry with speech recognition (see Pham et al., 2005). While the ‘agents’ in either of these examples can run on desktop computers, it is the fact that they are continuously running on the computer worn by the user that makes them interesting. The role of the wearable computer in these examples is to discretely run searches in the background and alert the wearer to interesting links and associations between the current topic of conversation (or typing) and information to which the computer has access. It might also be useful for the computer to track user activity and then to either record patterns of activity (in order to refine its model of context) or to offer information relevant to the user. Thus, a very common application domain for context-aware, wearable computer research is the museum visitor (Sarini & Strapparava, 1998). The idea is that when the visitor stands at a certain location, say in front of a painting, the computer offers information relating to the objects in that location. The manner in which the location (and other contextual factors) is used to select information and the manner in which the information is presented to the user might vary across applications. The examples considered thus far present the wearable computer as a device that is able to ‘push’ potentially relevant information to the wearer. However, rather than simply overloading the wearer with information, the role of the agents or context-awareness is to manage and tailor the information to the person.

Impact on User Performance

While there are many applications reported in the literature, there is a surprising lack of research into how effective these applications are in improving or otherwise changing user performance. One domain in which wearable computers have received both interest and support has been in maintenance (Mizell, 2003). A study exploring the use of head-mounted displays to support maintenance work show that performance could be improved, providing information was displayed in an appropriate format (Kancler et al., 1998). However, other studies have been more equivocal. Tasks requiring participants to follow instructions on a wearable computer or printed on paper have shown both the wearable computer (Baber et al., 1998) and the paper (Baber et al., 1999c; Ockerman et al., 1997; Siegel & Bauer, 1997) to lead to superior performance. One explanation of these differences lies in the design of the information that was presented, for example the manner in which information is presented can impact on overall performance times (Baber et al., 1999b; Sampson et al., 1993). One implication of these studies is that participants using the wearable computer tend to follow the same sequence of tests (as defined on the visual display), whereas
In terms of interacting with wearable computers and the appropriate devices to use, there has been very little work to date. While one might assume that the optimal interaction techniques would be ones that support hands-free interaction, such as speech recognition, studies suggest that walking has a negative impact on speech recognition performance (Oviatt, 2000; Price et al., 2004). In terms of entering data, Thomas et al. (1997) showed that a forearm mounted QWERTY keyboard led to superior performance over a five-button chording device or a virtual keyboard controlled using an isometric button. However, one might question the recommendation of a forearm mounted device, based on consideration of musculoskeletal strain. In terms of selecting objects on a display, Thomas et al. (1998) found that a touchpad mounted on the forearm was preferred by users, but that one mounted on the thigh lead to superior performance when sitting, kneeling or standing. Thus, the mounting of a pointing device can have a bearing on performance (although one might question whether pointing is an appropriate means of performing selection tasks on a wearable computer). Zucco et al. (2006) considered the performance of ‘drag and drop’ tasks while stationary and whilst walking, using different devices. They found that a gyroscopic mouse lead to best performance while stationary, but that touchpad or trackball were lead to better performance when the user was walking (and that all devices were superior to the Twiddler keypad).

Rather than seeing these studies as recommendations for specific interaction devices, I feel that they illustrate that the relationship between the activity that a person is performing and the demands of the ongoing activity in the environment interact in ways that require careful planning in the design of wearable computers. This brief discussion raises questions on one might develop standards for the deployment of wearable computers in these (and related) applications, and also what level of performance improvement one might expect from this technology.

SMART CLOTHING AND TEXTILES

It is worth mentioning the efforts to incorporate at least some aspects of technology into clothing and textiles (Berzowska, 2005). This could then take the form of clothing that has been adapted to incorporate some of the components (see Figure 6).

Alternatively, the textiles can exhibit some form of ‘smartness.’ Often this refers to “Textiles that are able to sense stimuli from the environment, to react to them and adapt to them...” (Van Langehoven & Hertleer, 2004). Taking these terms separately, we can ask how could textiles sense, react and adapt. In terms of sensing, there is a wide range of possible approaches, including thermal, chemical, mechanical, as well as

Figure 6. Incorporating consumer electronics into clothing—the Philips-Levi Strauss jacket

![Figure 6. Incorporating consumer electronics into clothing—the Philips-Levi Strauss jacket](image-url)
biosensors. Buechley (2006) shows how simple off-the-shelf sensors and actuators can be incorporated into items of clothing. In particular, this work, following the earlier work of Post and Orth (1997) demonstrates how fabrics can be knitted or woven to provide some electrical conductivity, and then fitted with components. In contrast, other researchers report ways in which it is possible to use to electrical properties of prepared fabrics, such as change in resistance, inductance or capacitance (Wijesiriwardana et al., 2003, 2004), to incorporate specific sensing capabilities, such as stretching (Farrington et al., 1999; Huang et al., 2006) or impact (Lind et al., 1997). What has yet to be fully realized from this smart textile work is that manner in which the user would interact with the ‘computing’ aspects in order to entry data or perceive displayed information. Thus, much of the research is still concerned with the development of the ‘bus’ onto which sensors, processors, displays, batteries, and so forth can be mounted.

**EMOTIONAL IMPACT OF WEARING COMPUTERS**

There are three broad categories of impact that will be considered in this section. The first concerns the emotional response to wearable computers by the people wearing these devices and the people with whom they interact. The second concerns the response of the computer to the emotions of the wearer, that is affective computing. The third concerns the manner in which collaboration can be supported within a group of wearable computer wearers.

While head-mounted displays have been used for many years in the domain of military aviation, they have yet to find widespread use on the ground. Most cinema goers will be familiar with concepts of wearable computers from science fiction films and might feel uncomfortable with ‘cyborgs’ in their midst. This discomfort could be particularly acute in stressful situations (as one of the paramedics in an earlier study pointed out to us, *If you’re just coming round from a heart attack, the last thing you’d expect to see if some robot headed bloke trying to take your pulse*). As there are so few commercial applications of wearable computers, these devices still represent something of a novelty and there has been very little research into how people might respond to people wearing such devices. One collection of anecdotal evidence can be found in Thad Starner’s Ph.D. Starner completed his Ph.D. at MIT, and with Brad Rhodes and Steve Mann developed a variety of wearable computers that they wore for prolonged periods of time. This meant that they would encounter members of the public on a regular basis and Starner recorded some of the reactions. A common response was to assume that the wearable computer was simply some form of familiar technology, perhaps a very sophisticated video camera or an advanced form of video game console or a medical device. The implication of this is that people might try to explain the technology using a familiar mental model but, as Starner points out, people have yet developed a mental model of a wearable computer on the basis of familiarity and exposure. While this is partly a matter of the unfamiliar appearance of the wearer it is also related to the manner in which the wearer interacts with the device. For example, Starner points out that when you ask someone the time they will raise their wrist in order to consult their watch, but with a head-mounted display one might simply glance up at the screen. This means that the physical behavioral cues might be far less obvious. This is, of course, similar to the way in which Bluetooth headsets allow people to speak on their mobile telephone in a manner that makes it looks as if they are speaking to themselves (indeed, some users of Bluetooth headsets make gestures that look as if they are holding a phone, for example holding the microphone or holding their hands near their faces). There will be a trade-off between the familiarity of wearable
computers, their appearance, and the ability of people to explain their use. However, there is an additional factor at play here (which is also hinted at by the Bluetooth headset discussion), and that is that the wearable computer provides information to an individual in a highly individualized manner—it is not possible for other people to see what the wearer is looking at or guess what they are doing. An analogy can be drawn in this instance with the person on a train reading a humorous book at laughing out loud—the other people feel uncomfortable because they can not read the what the person is reading. The removal of the observer from the source of information can be disconcerting and can lead to such comments as (from Starner) “we can’t tell if you’re talking about us behind our backs” or “when you wear your display, how can I tell if you are paying attention to me or reading your e-mail?”

One implication of the physical appearance and individualized interaction of wearable computers is the sense that people who wear such technology are different from ‘normal’ people. As Sheridan et al. (2000) note, people wearing computers could be perceived (by themselves or by the people around them) as a ‘cyborg community’ that is different from other people. With the use of communication and networking capabilities, it is possible for a group of wearable computer users to be able to share information and maintain contact as a community (Wellman, 2001). However, we should be careful to distinguish between the ability to maintain contact with other people (which one can do easily with mobile telephones, even to the extent of setting up talk groups) and the ability to share the wide range of information that wearable computing can support. This could simply mean the sharing of the same documents or video, but could also allow new forms of collaboration, sharing and exchange of information.

Context could be defined by changes in the physiological state of the wearer (Picard, 1997). This requires a more intimate means of recording data from the wearer, perhaps through monitoring of pulse or heart activity. The use of ‘context’ to initiate image capture has been demonstrated by several projects; most notably in Healey and Picard’s (1998) ‘StartleCam,’ in which changes in galvanic skin response (GSR) was used to trigger image capture. There has been surprisingly little attempt at extending this work in the years since it was reported, although over uses of ‘context’ in image capture have explored the use of ambient sound which is captured at the same time as the image (e.g., Frolich & Tallyn, 1999; Ljunblad et al., 2004). Bristow et al. (2005) used a set of context identifiers to take still images when ‘context’ changed, and showed that these were surprisingly consistent with photographs taken by humans. In addition to the computer ‘sensing’ the physiological responses of the wearer (and hence drawing some inference as to the affective state), it is also possible to infer the state of people with whom the wearer is interacting in order to develop ‘emotionally intelligent interfaces.’ By monitoring the changing facial expressions of ones conversational partner, it might be possible to provide support for people who find it difficult to judge the emotional response, for example people with Autism (El Kaliouby & Robinson, 2003).

**DISCUSSION**

Wearable computers continue to raise many significant challenges for human factors research. These challenges involve not only cognitive aspects of presenting information but also perceptual aspects of displaying the information against the backdrop of the everyday environment and physical aspects of mounting the devices on the person. This chapter has overviewed some of the developments in the field and offered some consideration of how these human factors can be considered. While the field is largely motivated by technological advances there is a need to carefully ground the developments in the physical and cognitive characteristics of the humans who are intended to wear them.
REFERENCES


Knight, J., Bristow, H.W., Anastopoulou, S., Baber, C., Schwitz, A., & Arvanitis, T.N. (2007). Uses of accelerometer data collected from a wearable...


Human Factors Problems of Wearable Computers


**KEY TERMS**

**Activity Models:** Predictive models of human activity, based on sensor data

**Augmentation Means:** Devices that can augment human behavior—a term coined by Doug Engelbart, and covering: **Tools & Artifacts:** the technologies that we use to work on the world which supplement, complement or extend our physical or cognitive abilities; **Praxis:** the accumulation and exploitation of skills relating to purposeful behavior in both work and everyday activity; **Language:** the manipulation and communication of concepts; **Adaptation:** the manner in which people could (or should) adapt their physical and cognitive activity to accommodate the demands of technology.

**Comfort:** Subjective response to wearing a wearable computer (ranging from physical loading to embarrassment)

**Context-Awareness:** The capability of a device to respond appropriately to changes in a person’s activity, environment, and so forth.

**Form-Factor:** The overall size (and shape) of a device

**Sensors:** Devices that produce digital output in response to some change in a measured parameter, for example dependent on environmental change or on user activity

**Wearable Computers:** Devices worn on the person that provided personalized, context-relevant information

**ENDNOTES**


2. Bainbridge (1987) argued that full automation can lead to the ironic situation that, the role of the human operator is to intervene when something goes wrong. However, the automation is such that the human is locked out of the process and has little understanding as to what is happening. Consequently, the human will not be able to intervene in an informed and efficient manner. Ultimately, it means that, by designing the human out of the system, the potential for a flexible and intelligent response to unknown situations is lost.
Human Factors Problems of Wearable Computers

3 http://www.bodymedia.com/main.jsp
4 http://www.extra.research.philips.com/pressmedia/pictures/wearelec.html

This work was previously published in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 158-175, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
ABSTRACT

In this chapter we discuss a number of recent studies that demonstrate the use of rational analysis (Anderson, 1990) and cognitive modelling methods to understand complex interactive behaviour involved in three tasks: (1) icon search, (2) graph reading, and (3) information retrieval on the World Wide Web (WWW). We describe the underlying theoretical assumptions of rational analysis and the adaptive control of thought-rational (ACT-R) cognitive architecture (Anderson & Lebiere, 1998), a theory of cognition that incorporates rational analysis in its mechanisms for learning and decision making. In presenting these studies we aim to show how such methods can be combined with eye movement data to provide detailed, highly constrained accounts of user performance that are grounded in psychological theory. We argue that the theoretical and technological developments that underpin these methods are now at a stage that the approach can be more broadly applied to other areas of Web use.

INTRODUCTION

With the rapid increase in Internet use over the past decade there is a growing need for those engaged in the design of Web technology to understand the human factors involved in Web-based interaction. Incorporating insights from cognitive science about the mechanisms, strengths, and limits of human perception and cognition can provide a number of benefits for Web practitioners. Knowledge about the various constraints on cognition, (e.g., limitations on working memory), patterns of strategy selection, or the effect of design...
Modelling Interactive Behaviour with a Rational Cognitive Architecture

decisions (e.g., icon style) on visual search, can inform the design and evaluation process and allow practitioners to develop technologies that are better suited to human abilities.

The application of cognitive psychology to human-computer interaction (HCI) issues has a long history going back to Card, Moran, and Newell’s (1983) introduction of the goals, operators, methods, and selection rules (GOMS) task analysis technique and model human processor (MHP) account of human information processing in the early 1980s. Since then, their cognitive engineering approach has developed into a family of methods (John & Kieras, 1994; Olson & Olson, 1990) which are widely used to produce quantitative models of user performance in interactive tasks.

Another, more recent approach to modelling human performance in interactive tasks has emerged in the last decade from theoretical and technological advances in research into cognitive architectures. Cognitive architectures are theories of the fundamental structures and processes that underlie all human cognition, of which there are several currently in existence including EPIC (executive process / interactive control; Kieras & Meyer, 1997), Soar (Laird, Newell, & Rosenbloom, 1987; Newell, 1990), and ACT-R (Anderson & Lebiere, 1998; Anderson et al., 2004). An important feature of these architectures is that they are all implemented as computer programming systems so that cognitive models may be specified, executed, and their outputs (e.g., error rates and response latencies) compared to human performance data.

Originally ACT-R and Soar were theories of central cognition only and did not explicitly specify mechanisms for perception or motor control. EPIC however, was unique in that from its inception it incorporated processors for cognition, perception, and motor control. Recent adaptations to ACT-R (Byrne & Anderson, 1998) and Soar (Chong & Laird, 1997) have now ensured that both architectures incorporate perceptual motor components that allow models to include visual attention processes and manual interactions with a keyboard and mouse. This is an important development for the study of HCI as cognitive models can now be embodied (Kieras & Meyer, 1997) in the sense that the architectures are now able to simulate perceptual-motor contact with computer interfaces and devices and so capture the complex interactions between the task environment, cognition, and perceptual-motor behaviour.

Modelling interactive behaviour with an embodied cognitive architecture has a number of advantages over the traditional cognitive engineering approach exemplified by GOMS and its relatives. Perhaps the most important of these is that computational models can actually execute the task, allowing a direct test of the sufficiency of the hypothesised processes. Second, although most cognitive architectures contain built-in timing parameters taken from the psychological literature, unlike cognitive engineering models, they do not require prior estimated times for all subcomponents of a task. In addition, some architectures — such as ACT-R and Soar — contain learning mechanisms which allow them to model various effects of practice on performance. This allows cognitive architectures to be used to model novel tasks, novice users, or tasks involving components without prior time estimates.

One of the promises of embodied cognitive architectures is that, once they are equipped with sufficient knowledge, they will begin to provide a priori predictions of user performance and eventually evolve into artificial users that can be employed to evaluate novel tasks and environments (Ritter, Baxter, Jones, & Young, 2000; Young, Green, & Simon, 1989). In this chapter we will describe one of these architectures, ACT-R, and show how it has been used to provide detailed and sophisticated process models of human performance in interactive tasks with complex interfaces. ACT-R is an appropriate choice for this discussion because, in contrast to other cognitive architectures, ACT-R also embodies
the rational theory of cognition (Anderson, 1990) which analyses cognitive phenomena in terms of how they are adapted to the statistical structure of the environment. Rational analysis and ACT-R’s mechanisms have been used recently to provide novel insights into Web-based interactions. The chapter proceeds as follows: First we describe the basic assumptions and mechanisms of rational analysis and the ACT-R cognitive architecture. We then show how these have been used to develop a model of information foraging on the Web and discuss the model in relation to a rational analysis model of the task and the data from eye-tracking studies of interactive search. In the final sections of this chapter we briefly outline ACT-R models of two interactive tasks; graph reading (Peebles & Cheng, 2003) and icon search (Fleetwood & Byrne, in press). Although neither of these studies involves a specifically Web-based task, they both describe user interaction with items commonly found on Web pages. They are also illustrative of a methodology that combines task analysis, eye tracking, and formal modelling to provide a detailed account of the cognitive, perceptual, and motor processes involved in the performance of the task. These studies are also useful because in both cases the model is validated by comparing the simulated eye movements with those recorded from human subjects. Both studies, therefore, are clear demonstrations of a novel approach to understanding interactive behaviour that can be applied to Web-based tasks.

RATIONAL ANALYSIS

Rational analysis (Anderson, 1990) is a method for understanding the task an agent attempts to complete. It assumes that humans have evolved cognitive mechanisms that are useful for completing tasks that we encounter in our environment, and that these mechanisms work in an efficient way to complete these tasks. Therefore, rather than concerning ourselves with firstly trying to define the cognitive mechanisms required by the agent to solve the task, rational analysis suggests that we should consider the structure of the task itself, the environment in which it is encountered, together with some minimal assumptions about the computational limitations of the system. From these initial statements the analysis proceeds by the specification of an optimal solution to the problem and the comparison of human behavioural data to see how close an approximation it is to the optimal solution.

By identifying the best way to complete the task (the optimal strategy) we can often infer what the cognitive mechanisms of a rational agent must be as although humans do not always complete tasks in the most optimal way their behaviour is usually similar to the optimal strategy. That is, humans usually behave in such a way that they appear to be trying to complete their tasks in the most efficient manner, that is, they try to maximise their returns while minimising the cost of achieving their goals.

Rational analysis has been applied to several aspects of human cognition (see e.g., Oaksford & Chater, 1998), from the original analyses of memory, categorisation, causal inference, and decision making conducted by Anderson (1990), to more recent analyses of exploratory choice (Cox & Young 2004; Young, 1998) and the updating of memory during tasks in dynamic environments (Neth, Sims, Veksler, & Gray, 2004).

THE ACT-R COGNITIVE ARCHITECTURE

ACT-R is a theory of human cognition developed over a period of 30 years by John Anderson and his colleagues (Anderson & Lebiere, 1998; Anderson et al., 2004) that incorporates the theory of rational analysis. It is a principal effort in the attempt to develop a unified theory of cognition (Newell, 1990). As a cognitive architecture, ACT-R attempts to specify the basic cognitive
structures and processes that underlie all human cognition.

Figure 1 illustrates the components of the architecture relevant to our discussion. ACT-R consists of a set of independent modules that acquire information from the environment, process information, and execute motor actions in the furtherance of particular goals. There are four modules that comprise the central cognitive components of ACT-R. Two of these are memory stores for two types of knowledge: a declarative memory module that stores factual knowledge about the domain, and a procedural memory module that stores the system’s knowledge about how tasks are performed. The former consists of a network of knowledge chunks whereas the latter is a set of productions, rules of the form “IF <condition> THEN <action>”: the condition specifying the state of the system that must exist for the rule to apply and the action specifying the actions to be taken should this occur. The other two cognitive modules represent information related to the execution of tasks. The first is a control state module that keeps track of the intentions of the system during problem solving, and the second is a problem state module that maintains the current state of the task.

In addition to these cognitive modules there are four perceptual-motor modules for speech, audition, visual, and motor processing (only the latter two are shown in Figure 1). The speech and audition modules are the least well-developed and, at present, simply provide ACT-R with the capacity to simulate basic audio perception and vocal output for the purpose of modelling typical psychology experiments. The visual and motor modules are more well-developed and provide ACT-R with the ability to simulate visual attention shifts to objects on a computer display and manual interactions with a computer keyboard and mouse.

Each of ACT-R’s modules has an associated buffer that can hold only one chunk of information from its module at a time, and the contents of all of the buffers constitute the state of an ACT-R model at any one time. Cognition proceeds via a pattern matching process that attempts to find productions with conditions that match the current contents of the buffers. There then follows a process to select the “best” production from those that match the conditions, after which the most appropriate production “fires” and the actions (visual or manual movements, requests for the retrieval of a knowledge chunk from declarative memory, or modifications to buffers) are performed. Then the matching process continues on the updated contents of the buffers so that tasks are performed through a succession of production rulefirings. As an example, two production rules (written in English rather than in ACT-R code) that instantiate part of a search task may look something like Box 1.

The processing in ACT-R’s modules is serial but the modules run in parallel with each other so

**Box 1.**

```
IF the goal is to find the meaning of “eudaimonia” (control state)
AND there is nothing in declarative memory about “eudaimonia” (declarative)
THEN set the goal to search the WWW for “eudaimonia” (control state)

IF the goal is to search the WWW for “eudaimonia” (control state)
AND the Web browser is open (problem state)
THEN look for the menu labelled “Bookmarks” (visual)
AND update the problem state to “looking for Google” (problem state)
```
that the system can move visual attention while also moving the mouse and attempting to retrieve knowledge from declarative memory. ACT-R processes also have associated latency parameters taken from the psychology literature. For example, it typically takes 50 ms for a production to fire and the time taken to move the mouse cursor to an object on the computer screen is calculated using Fitts’ Law (Fitts, 1954).

ACT-R implements rational analysis in two ways. The first is its mechanism for retrieving knowledge chunks from declarative memory which is based on the notion of activation. Each chunk in declarative memory has a level of activation which determines its probability and latency of retrieval, and the level of activation for a chunk reflects the recency and frequency of its use. This enables us to understand how rehearsal of items in a short-term memory task can boost the activation levels of these chunks and consequently increase the chances of recall/retrieval from declarative memory. The level of activation of a chunk falls gradually over time, and without retrieval or activation spreading from chunks in the current goal, it may fall below a threshold level which then results in retrieval failure. This enables ACT-R models to forget knowledge without having to explicitly delete chunks from the declarative memory store.

The second way that ACT-R implements rational analysis is in its mechanism for choosing between alternative production rules. According to rational analysis, people choose between a number of options to maximise their expected utility. Each option (i.e., production rule) has an expected probability of achieving the goal and an expected cost. It is assumed that when carrying out computer-based tasks people interact with the task environment and choose actions that will optimise their efficiency (i.e., maximise the probability of achieving the goal while minimising the cost, usually measured in units of time). At each decision step in the cycle, therefore, all possible production rules that match against the current goal are proposed in a choice set, and the one with the highest level of efficiency is chosen and executed.

ACT-R has been used to model a wide range of cognitive phenomena (Anderson & Lebiere, 1998), and in recent years, with the inclusion of the perceptual-motor modules, it has been applied to a number of complex interactive tasks in the area of HCI and human factors research, for example, menu selection (Byrne, 2001), cell phone menu
interaction (St. Amant, Horton, & Ritter, 2004), and driving (Salvucci & Macuga, 2002). Although individually these models do not yet offer us a virtual “user” which can be sat in front of a Web browser and asked to complete any goal, together they provide us with insights into how and why users behave in particular ways, for example, when searching for information on the Web. In this chapter we will concentrate on three particular areas of work that are relevant to understanding Web behaviour: icon search, graph reading, and information foraging on the WWW.

MODELLING INTERACTIVE BEHAVIOUR

In the following section, we will summarise a number of recent studies which employ rational analysis, cognitive modelling, eye tracking, or a combination of all three, to understand human performance in Web-based or HCI tasks. We first discuss recent efforts to model information foraging and interactive search on the WWW. These studies show how ACT-R and rational analysis can be successfully applied to explain different aspects of people’s behaviour when conducting interactive search tasks. This can include both high-level behaviours such as backtracking through Web-pages and low-level behaviours such as patterns of visual attention obtained from eye-tracking studies. We then describe two studies which combine experimental data collection, eye movement recording, and cognitive modelling methods using ACT-R to provide detailed accounts of the cognitive, perceptual, and motor processes involved in the tasks. These studies were chosen because both develop a detailed process model which not only captures the human response time data from the experiment, but also provides a close match to the patterns of visual attention revealed by the eye movement study. This level of detail in modelling is still relatively uncommon and the strong constraints added by seeking to match model and human eye movement scan paths during the course of the task provide a further validation of the models.

Information Foraging on the World Wide Web

Information foraging theory (IFT; Pirolli & Card, 1999; Pirolli, 2005) describes an account of information gathering behaviour based on the ecological behaviours of animals when foraging for food. The account can be applied to situations in which people are searching for information in a number of different situations such as in a library or on the WWW. The theory rests on rational analysis in that it proposes that human behaviour is directed by the objective to maximise gain and minimise effort, and that this process is sensitive to changes in the environment. In contrast to animal studies, where the assumption is that animals seek to reduce the ratio of calorie intake to energy expenditure, the assumption in IFT is that people attempt to reduce the ratio of information gained to time spent.

The way in which the environment is structured determines the costs of search for information. For example, the structure of a Web site will determine how many pages the user has to navigate through in order to satisfy his/her goal. When searching for information on the WWW, many people make use of search engines. After entering some key words the user is presented with a list of search results which are usually ordered in terms of their relevance to the key words. Each of the results returned can be considered to be a “patch” of information. The user has to choose to either investigate one of the patches or to redefine their search criteria. Conducting another search using different key words will result in a change in the environment. This process is known as enriching the environment as it is hoped that the result is that the cost of obtaining the required information will be reduced compared to the perceived cost of obtaining it in the previous environment. Deci-
sions about whether or not to pursue a particular information patch or to continue enriching the environment are based on a number of factors such as the perceived value of the information returned, the perceived costs of acquiring that information, interface constraints, and previous knowledge.

The decision to forage within a particular patch of information is based on an ongoing assessment of information *scent*. Information scent is the perception of the value of the distal information based on the proximal information available, that is, it is an estimate of the relevance of the information contained on a yet unseen page based on the cues from the icon or wording of the link on the page currently viewed. The theory predicts that as more time is allocated to within-patch foraging, the rate of information return increases but only up to an optimal point, after which the rate starts to decrease. Therefore, after a particular amount of within-patch foraging (searching within a Web site) it becomes more profitable to move to the next patch (select another Web site from the list of search results) even though there are still pages within the previous patch that have not yet been visited.

**SNIF-ACT**

Scent-based Navigation and Information Foraging in the ACT architecture (SNIF-ACT) (Pirolli & Fu, 2003) is a model of human behaviour in an interactive search task. The model makes use of ACT-R’s spreading activation mechanism so that the information scent of the currently viewed Web page activates chunks in declarative memory as does the spreading activation from the goal. Where these two sources of activation coincide there are higher levels of activation and this indicates a high degree of relevance between the goal and the page being attended to. This activation is what ultimately drives the behaviour of the model. The model includes the use of search engines to provide a set of search results and the processing of the page that is returned. The links on the page are attended to and eventually one of the links is selected.

The behaviour of the model is compared to user behaviour and successfully demonstrates that people tend to select the highest scent item in a list. SNIF-ACT does this by assessing the information scent of all the links on a page and then choosing the highest one. The model is also able to explain the point at which a user abandons a particular Web site and returns to the search results in order to select another item from the list or selects a link that takes them to another Web site. If the mean information scent of the currently viewed page is lower than the mean information scent of a page on another site the model selects that action that takes them to the other site.

**Eye-Tracking Experiments in Interactive Search**

When presented with a list of search results or items on a menu within a Web site (i.e., a patch of information), the user has to choose between selecting an item which will move him/her to another patch and doing some assessment on either the currently attended item or some other item in the list (i.e., consume the information presented within the current patch). As has been mentioned previously, IFT proposes that the user will make use of the information scent of the items to guide their behaviour. If the information scent of a particular item in the list is higher than the rest (i.e., that item appears to be relevant to the task and the user believes that clicking it will lead them to better information) then the item will be selected.

Eye-tracking experiments have been used to investigate what people attend to when conducting interactive search tasks (Brumby & Howes, 2004; Silva & Cox, 2005). Participants were given an information goal and a list of items and asked to select the label that they thought would lead to the
Modelling Interactive Behaviour with a Rational Cognitive Architecture

Brumby and Howes demonstrated that people often examine only a subset of the list before selecting the target item, and that this behaviour is affected by the relevance of the other items in the list. When the other items in the list are more relevant to the goal (i.e., they have high levels of information scent), people tend to look at more items in the list and also tend to look at individual items on more occasions than when the items are irrelevant. When there are a number of items with high scent (i.e., two or more items look like they would lead to relevant information) people need to consider more items than when only one item looks sensible.

However, one limitation of this work is that the analysis of eye-tracking data is rarely sensitive enough to determine whether a lack of fixation of the eyes on an item really means that people have not assessed the relevance of the item. In order to address this, Silva and Cox (2005) additionally employed a recognition task in their study in order to assess the level of processing of each item in the list.

Figure 2 represents a simplified scan path of a participant completing one of these tasks. The items are represented on the y axis with time along the x axis. The highlighted item is the target item and was selected by the participant. The figure demonstrates how the user starts at the top of the list and scans down the list fixating items in the list. Some of the items (3 & 6) are skipped over. The results from Silva and Cox’s (2005) recognition task suggest that in such cases the lack of fixations of particular items in the menu can be explained by parafoveal processing. However, parafoveal processing can only explain lack of fixations on up to two items below the last fixation (i.e., items 8 & 9) and cannot explain why the user does not attend to other items in the list (i.e., items 10 to 16).

SNIF-ACT would be able to produce a trace that would match the behaviour of users in these studies in terms of which items from the menus the user selected. However, the model does not account for the fact that some of the items in the menus were not assessed by the users as it assumes that users have knowledge about information scent of all the items in the list and then selects the item with the highest level of scent. Consequently, SNIF-ACT is unable to provide us with any explanation for why users should choose to select an item when they have not even read the entire list presented to them.

Cox and Young (2004) propose an alternative model to that of SNIF-ACT that is able to capture this fine-grained level of detail of user behaviour. Their model is a rational analysis of an interactive search task that provides a rational explanation of why the user would select an item without first assessing all the items in the list.

In interactive search, the agent has the goal of selecting the item that will lead to goal completion. However, as the menu presented is novel, the first thing that the model has to do is to gain some information about the menu. The model

**Figure 2. A simplified scan path of a participant performing an interactive search task**
therefore includes two types of exploratory acts (EAs) (these are the different types of things the model can do): assess information SCENT and ANTICIPATE the result of selecting this item. The SCENT EA should be thought of as being an amalgamation of perceiving the label, reading the label (at a lexical level), and considering the semantic similarity between the label and the current task. The ANTICIPATE EA should be thought of as some additional cognitive effort that considers whether the label is likely to lead to the goal. For example, given the goal of finding an armchair for your living room on a furniture shop Web site, imagine the model considering the first item in the menu “home.” The SCENT EA would return a moderately high rating as the label has a moderately high level of information scent given the goal (“home” and “armchair”). The ANTICIPATE EA models the agent’s consideration of whether the label home is likely to lead to the home page of the site, or to a list of home furnishings. Each of these EA types has a cost associated with it with the ANTICIPATE EA type being more expensive in mental effort than the first type. There is also a fixed cost of moving attention from one item in the menu to the next.

Before assessing any items, the model “knows” the number of items in the menu and considers each of these items to be equally (ir)relevant to completing the task. The scent ratings of the items in the menu are used as the basis for determining the new relevance (R) value of an item following an assessment. On each page, the set of relevancies R_i are mapped into a set of probabilities P_i by the transformation P_i = odds(R_i)/∑odds(R_j), where odds(R) is defined in the standard way as odds(R) = R/(1–R). Note that ∑P_i = 1, reflecting the fact that exactly one option on the page leads to the goal.

When the model is run on a set of menus it demonstrates how different patterns of information scent result in different behaviours. As Brumby and Howes (2004) demonstrated, the levels of information scent of both the goal item and the distractors affect behaviour. However, it is also interesting to note that the model predicts that just the change in position of the goal item relevant to the distractors results in different patterns of behaviour: Sometimes the model predicts that users will scan to the bottom of the menu before selecting the target item, and other times they will select the item immediately after assessing the item leaving other items in the menu unassessed. To explain how this occurs we will compare the behaviour of the model when the high scent item is in position two (as an example of occurring early in the menu) and in position 12 (as an example of occurring late in the menu) in more detail. In both examples, initially, all 16 menu items are rated equally and all have an R value of 0.06. The relevance values are translated into efficiencies (E) which are then used to determine which of the EAs is most likely to lead to the goal and therefore which EA is executed in each cycle. In the first cycle, the EA that proposes assessing the scent of the first item in the menu is rated as having the highest E value due to it having the lowest cost. Consequently, the model assesses the first item which gets rated as very low scent. As a result, the new R value of this item is set at 0. On the next cycle, the EA that proposes SCENT assessment on the second item in the list is the most efficient (due to the lower cost) so this item gets assessed. This behaviour continues until the model assesses the high scent item.

In menus where the high scent item occurs early on in the menu, the second item in the menu gets an R value of 0.5097 which raises the probability that this item will lead to the goal to 0.6220. On the following cycle the R value of the high scent item leads to an E value of 0.008 while the second best item (an item yet to be assessed) has an R value of 0.06 which results in an E value of 0.006. Although the E values of the two EAs are very similar, one is larger than the other, and this is what determines which EA is chosen.
In our example of a menu where the high scent item occurs later on in the menu, the relevance of each of the low scent items that have already been assessed falls to 0. When the model assesses the twelfth item its R value is 0.5097, which raises the probability that this item will lead to the goal to 0.6220. On the following cycle the R value of the high scent item only has an E value of 0.005 while the item with the best efficiency (an item yet to be assessed) has an R value of 0.05 which results in an E value of 0.006. The result is that the model continues to assess each item in the menu until it reaches the bottom because the efficiency of conducting a SCENT assessment of a new item is greater than the efficiency of conducting the ANTICIPATE assessment on the high scent item in position 12. This has the effect of slowly increasing the probability of the item in position 12 leading to the goal.

The detail of the model explains that the reason the behaviour is different for the two types of menus is because the detail of the mathematics of the rational analysis. Comparisons of the traces of the model with the empirical data suggest that the model provides a good explanation of the cognitive processes involved in this task. This suggests that participants make an assessment of the relevance of a label to the current goal and then, together with the estimated relevance of previous items, choose to either (1) select that item as the one that will lead to the goal, (2) conduct some further assessment of the current item, or (3) move on to another item and assess that. Which of these EAs is chosen is driven by the pattern of information scent that has been experienced so far.

The model provides us with an explanation of how and why the position of the goal and the quality of the distractor items affect the behaviour of the participants on the task. Regardless of the pattern of scent of the menu, the model predicts that the agent will tend to stop exploring the menu as soon as it comes across a menu item that has high information scent (self-terminates) if this is encountered early in the menu. On menus where there is one high scent item among a set of low scent items and the high scent item occurs later in the menu, the agent continues to assess the other items in the menu before conducting further assessment of the high scent item and finally selecting it. The model enables us to explain why we see these different patterns of behaviour on menus which have such similar patterns of information scent. This is due to the effect of the interdependence of the probability that each of the items will lead to the goal. The actual point on the menu at which the model swaps from one behaviour to the other is sensitive to a number of factors such as the length of the menu and the costs of the EAs. It would appear therefore that it is in the nature of interactive search that there are close calls which suggest that people can rationally do either behaviour and that a number of factors have an effect on the behaviour of participants exploring real menus.

Together the two models described previously provide us with a good understanding of how people perform search tasks on the WWW. SNIF-ACT and the rational model explain different aspects of the interaction: SNIF-ACT demonstrates the higher level, page by page, link following behaviour seen in such tasks, whereas the rational model explains the lower level interactions with just one page. Given information about the information scent of the items on a new Web site both models are able to make predictions about user behaviour on the site.

**Modelling Graph Reading**

Peebles and Cheng (2003) conducted an experiment, eye movement study and cognitive modelling analysis to investigate the cognitive, perceptual, and motor processes involved in a common graph-reading task using two different types of Cartesian graph. The purpose of the study was to determine how graph users’ ability to retrieve information can be affected by presenting the same information in slightly different types of the same
class of diagram. The two types of graph, shown in Figure 3, represent amounts of UK oil and gas production over two decades. The only difference between the two graph types is in which variables are represented on the axes and which are plotted. In the Function graphs, the argument variable (AV: time in years) is represented on the x-axis and the quantity variables (QV: oil and gas) on the y-axis whereas in the Parametric graphs, the quantity variables are represented on the x and y axes and time is plotted on the curve.

In the experiment, participants were presented with the value of a “given” variable and required to use the graph to find the corresponding value of a “target” variable, for example, “when the value of oil is 2, what is the value of gas?” This type of task has typically been analysed in terms of the minimum sequence of saccades and fixations required to reach the location of the given variable’s value and then from there to the location of the corresponding value of the target variable (Lohse, 1993; Peebles & Cheng, 2001, 2002; Peebles, Cheng, & Shadbolt, 1999). Experiment participants (some of whom had their eye movements recorded) completed 120 trials, each participant using only one graph type. The 120 questions were coded into three classes (QV–QV, QV–AV, and AV–QV) according to which variable’s value was given and which was required (QV denotes a quantity variable, oil or gas, and AV denotes the argument variable, time). On each trial, a question (e.g., “GAS = 6, OIL = ?”) was presented above the graph and participants were required to read the question, find the answer using the graph on the screen and then enter their answer by clicking on a button labelled Answer in the top right corner of the window which revealed a circle of buttons containing the digits 0 to 9. RTs were recorded from the onset of a question to the mouse click on the Answer button.

Figure 3. Function and parametric graphs used in Peebles and Cheng (2003) depicting values of oil and gas production for each year

Notes: The graphs on the left (labelled 1) show years 1970 to 1979 while those on the right (labelled 2) show years 1980 to 1989. Dashed lines indicate the optimal scan path required to answer the question, “when the value of oil is 3, what is the value of gas?”
the screen and then enter their answer by clicking on a button labelled Answer in the top right corner of the window which revealed a circle of buttons containing the digits 0 to 9. RTs were recorded from the onset of a question to the mouse click on the Answer button.

The RT data from the experiment, displayed in Figure 4, showed that the graph used and the type of question asked both had a significant effect on the time it took for participants to retrieve the answer. This was all the more surprising because, for two of the three question types, participants were faster using the less familiar parametric graphs by nearly a second.

The results of the eye movement study were also surprising. It was found that in 63% of trials (irrespective of the graph used or question type being attempted), after having read the question at the start of a trial, participants redirected their visual attention to elements of the question at least once during the process of problem solving with the graph. This was not predicted by the simple minimal fixation sequence account outlined previously but two possible explanations may be provided: (1) participants initially encode the three question elements but are unable to retain all of them in working memory and retrieve them by the time they are required to do so, or (2) to reduce the probability of retrieval failure, participants break the problem into two sections, the first allowing them to reach the given location and the second to then proceed to the target location corresponding to the solution.

Peebles and Cheng (2003) constructed two ACT-R models of the experiment (one for each graph type) that were able to interact with an exact replica of the experiment software. The models consisted of a set of productions to carry out the six basic subgoals in the task; (1) read the question; (2) identify the start location determined by the given variable; (3) identify the given location on the graph representing the given value of given variable; (4) from the given location, identify the target location representing the required variable; (5) identify the target value at the target location; and (6) enter the answer. Many of the productions were shared by the two models, the main difference between them being the control structure that sequences the execution of the productions. Figure 4 shows that the mean RTs from the parametric and function graph models are a good fit to the observed data ($R^2 = .868$, RMSE = 0.123, and $R^2 = .664$, RMSE = 0.199 respectively). Perhaps more importantly however, were the insights into the observed eye movement data that came from the modelling process itself. When ACT-R focuses attention on an object on the screen, representations of the object and its location are created in

Figure 4. Mean response times for experimental participants and ACT-R models for each question type (Peebles & Cheng, 2003)
the system’s visual buffers which can be accessed by productions. Eventually these representations go into declarative memory with initial activation values and, as long as these values are above a certain threshold, they can be retrieved by the cognitive system and replaced in a buffer. However, ACT-R includes a mechanism by which the activation of representations in declarative memory decreases over time which allows it to simulate processes involved in forgetting. These mechanisms played a crucial role in the ACT-R models’ ability to capture the eye movement data observed in the experiment. At the start of each trial, the models read the three question elements and during the problem solving these elements are placed in declarative memory. As a consequence, at least one question element must be retrieved from memory at each stage of the problem in order to continue. However, as soon as a question element is placed in declarative memory its activation starts to decay and, as a consequence, the probability that it cannot be retrieved increases. Typically, if a retrieval failure occurs, an ACT-R model will halt as it does not have the appropriate information to solve the problem. During the process of model development it was found that on a significant proportion of trials the model was not able to retrieve question elements at the later stages of the trial because their activation had fallen below the retrieval threshold. As a consequence new productions had to be added to allow the model to redirect attention to the question in order to re-encode the element and then return to solving the problem. This was precisely the behaviour observed in the eye movement study. This is illustrated in Figure 5 which compares screen shots of the model scan path and eye movements recorded from one participant for the same question using the 1980’s parametric graph. The numbered circles on the model screen shot indicate the sequence of fixations produced by the model. The pattern of fixations in both screenshots is remarkably similar.

Modelling Icon Search

Fleetwood and Byrne’s study of icon search (2002, in press) is another demonstration of how an ACT-R cognitive model can provide a detailed account of the cognitive and perceptual processes involved in a common HCI task that closely matches people’s response times (RTs) and

![Figure 5. Screen shots showing an experimental participant’s eye movement data (left) and the ACT-R model’s visual attention scan path (right) for the QV–QV question “oil = 6, gas = ?” using the 1980's parametric graph.](image)

Note: In the model screen shot, numbered circles on the scan path indicate the location and sequence of fixations.
patterns of eye movements. Fleetwood and Byrne’s model differs from that of Peebles and Cheng (2003) in that it incorporates eye movements and movement of attention (EMMA) (Salvucci, 2001), a computational model of the relationship between eye movements and visual attention. EMMA can be easily integrated into the ACT-R architecture, allowing models to make more detailed predictions of actual eye movements, rather than simple shifts of visual attention.

One of the main aims of Fleetwood and Byrne’s research is to investigate the notion of icon “quality” (defined in terms of an icon’s distinctiveness and visual complexity) and to examine the effect that differences in quality may have on identification performance. They created three classes of icon (examples of which are shown in Figure 6). “Good” quality icons were designed to be easily distinguishable from others based on the primitive features of colour and shape. All icons in this set were a combination of one colour (from six) and one shape (from two).

In contrast, “poor” quality icons were designed to be distinguishable only by a relatively careful inspection but to be relatively indistinguishable in a large distractor set. These poor quality icons were all of the same basic shape and colour (a combination of black, white, and shades of grey). An intermediate class of “fair” quality icons was also designed with shapes more distinctive than the poor quality icons but more complex than the good quality icons, and with the same range of greyscale colours as the poor quality icons. The main effect of the manipulation was to produce a different similarity structure for each class of icons. Good quality icons could be identified as a single combination of features, for example, “yellow triangle.” In contrast, fair quality icons were defined by more than one combination of features (typically three, for example: “grey rectangle; black square; black diagonal-right”), some of which were shared with other icons. In the poor quality group, icons were defined by an average of four feature combinations and many more of these were shared by several other icons in the group. From the visual search literature, it can be predicted that search time will increase as icon distinctiveness decreases. An additional factor in Fleetwood and Byrne’s (2006) study also known to affect search time (at least for certain stimuli) is the number of distractors in the display, with search time increasing with the number of distractors in the search set. In their experiment, Fleetwood and Byrne had search sets of 6, 12, 18 and 24 icons.

In the experiment, participants were required to find, as rapidly as possible, different quality target icons in search sets of differing sizes. On each trial, a target icon and file name were presented followed 1500 ms later by a button labelled Ready for the participant to click when he/she felt ready to continue. When this button was clicked, the target icon was replaced by the search set and the participant had simply to look for the target icon and click on it as quickly as possible; when an icon was clicked upon, the next trial started. Participants completed a total of 144 trials, involving all levels of the search set and icon quality variables, and on each trial the participant’s RT (the duration between clicks on the Ready button and an icon in the search set) was recorded. The results of the experiment (shown in Figure 7) revealed that, as predicted, both icon quality and search set size had a significant effect on search time.

Figure 6. Examples of icons of good, fair, and poor quality used in the experiment of Fleetwood and Byrne (in press)
To provide an explanation of their data, Fleetwood and Byrne (2006) produced an ACT-R model of the task that was able to interact with the same experiment software as the participants. As described previously, each experiment trial is comprised of two stages, the first where the target icon and its file name are encoded and the second in which it is sought. The model has a set of seven productions to carry out the first stage: (1) locate the target icon and (2) encode an attribute pair (e.g., "grey rectangle"), (3) look below the icon and (4) encode the associated file name, and finally (5) locate and (6) click on the “Ready” button. In the second stage, the model locates and attends to an icon with the previously encoded target feature and then shifts visual attention to the file name below it. If the file name matches the target file name, visual attention is returned to the icon and the mouse clicks on it. If the file name is not the target, however, the model continues the search by locating another icon at random with the same target features. This sequence of events requires four productions and takes 285 ms to complete.

Figure 7 reveals a close correspondence between the mean RTs produced by the model and those of the experiment participants ($R^2 = .98$, RMSE = 126ms) and shows that an ACT-R model based on the similarity structure of the search set and the strategy of identifying a single combination of features and random search can provide a reasonable account of the data. However, Byrne, Anderson, Douglass, and Matessa (1999) had shown in an earlier study of visual search in a menu selection task that alternative strategies can produce similar aggregate RTs, necessitating the incorporation of eye movement data to add further constraints on the proposed theory. As a result, Fleetwood and Byrne (2006) carried out an eye movement study to test their model further and found two major discrepancies between the observed eye movements and the patterns of visual attention produced by their model. First, they found that, although the model successfully reproduced the patterns of visual attention across the icon quality and set size conditions, for all conditions the number of saccades per trial produced by the model was significantly greater than those recorded in the experiment. Second, when analysing the eye movement data, Fleetwood and Byrne found that patterns of icon search were not random as their model predicted, but were systematic, in the sense that participants sought to minimise the distance between successive fixations, typically looking at target icons closest

---

**Figure 7. Response time by set size and icon quality for Fleetwood and Byrne's (in press) revised model and the experiment data**
to their current fixation point. This produced a search pattern that revealed a systematic scanning of areas of the display.

Both of the discrepancies between the model and human data are explained by Salvucci’s (2001) EMMA model. It has been demonstrated previously that the relationship between eye movements and visual attention is not direct, and that people often do not move their eyes to their focus of attention (e.g., Henderson, 1992; Rayner, 1995). EMMA attempts to capture this relationship by providing an account of if and when eye movements occur, and if they do occur, the location of their landing relative to their targets. Integrating EMMA into ACT-R allows models to simulate actual eye movements rather than just visual attention shifts and provides a more realistic output to be compared with human eye movement data. In addition, EMMA predicts that efficient search strategies minimise average saccade distance, resulting in search patterns in which objects nearest to the current fixation point are examined soonest.

Fleetwood and Byrne (2006) modified their model’s search strategy according to the EMMA account and incorporated EMMA’s eye movement computations into their model, resulting in a greatly improved fit (shown in Figure 8) to the human eye movement data ($R^2 = .99$, RMSE = 0.58).

CONCLUSION

In this chapter we have presented a number of recent examples of research that we believe clearly demonstrate the value of rational analysis and cognitive modelling in the study of complex interactive behaviour. Such tasks typically involve the complex interaction of three elements: (1) the perceptual and cognitive abilities of the user; (2) the visual and statistical properties of the task environment; and (3) the specific requirements of the task being carried out. The use of rational analysis and an embodied cognitive architecture such as ACT-R allows all three of these elements to be brought together in an integrated theoretical account of user behaviour. Rational analysis provides a set of assumptions and methods that allow researchers to understand user behaviour in terms of the statistical structure of the task environment and the user’s goal of optimising (i.e., reducing the cost/benefit ratio of) the interaction. Developing

Figure 8. Mean number of shifts of visual attention per trial made by Fleetwood and Byrne’s (in press) revised model relative to the mean number of gazes per trial made by participants
cognitive models of interactive behaviour in a cognitive architecture such as ACT-R allows researchers to specify precisely the cognitive factors (e.g., domain knowledge, problem-solving strategies, and working memory capacity) involved. In addition, the recent incorporation of perceptual-motor modules to cognitive architectures allows them to make predictions about users’ eye movements during the entire performance of the task, which can be compared to observed eye movement data — a highly stringent test of the sufficiency and efficacy of a model. The use of these methods has increased rapidly over the last 5 years, as has the range of task interfaces being studied. Although we are still a long way from achieving the goal of an artificial user that can be applied “off the shelf” to novel tasks and environments, the models of interactive behaviour described here demonstrate a level of sophistication and rigour still relatively rare in HCI research. As these examples illustrate, developing more detailed accounts of interactive behaviour can provide genuine insights into the complex interplay of factors that affect the use of computer and Web technologies, which may inform the design of systems more adapted to their users.

NOTE

All correspondence to: David Peebles, Department of Behavioural Sciences, University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, UK; D.Peebles@hud.ac.uk.

REFERENCES


*This work was previously published in Human Computer Interaction Research in Web Design and Evaluation, edited by P. Zaphiris, pp. 290-309, copyright 2007 by Information Science Publishing (an imprint of IGI Global).*
Chapter 3.11
Device Localization in Ubiquitous Computing Environments

Rui Huang
University of Texas at Arlington, USA

Gergely V. Záruba
University of Texas at Arlington, USA

Sajal Das
University of Texas at Arlington, USA

ABSTRACT

In this chapter, we will study the localization problem in ubiquitous computing environments. In general, localization refers to the problem of obtaining (semi-) accurate physical location of the devices in a dynamic environment in which only a small subset of the devices know their exact location. Using localization techniques, other devices can indirectly derive their own location by means of some measurement data such as distance and angle to their neighbors. Localization is now regarded as an enabling technology for ubiquitous computing environments because it can substantially increase the performance of other fundamental tasks such as routing, energy conservation, and network security. Localization is also a difficult problem because it is computationally intractable. Furthermore, it has to be implemented in a highly dynamic and distributed environment in which measurement data is often subject to noise. In this chapter, we will give an overview of localization in terms of its common applications, its hardware capacities, its algorithms, and its computational complexity.

INTRODUCTION

In a ubiquitous computing environment, devices are often connected to one another on the fly to form an infrastructure-less network that is fre-
Device Localization in Ubiquitous Computing Environments

quenty referred to as a mobile ad hoc network (MANET). Since MANET serves as an abstract model that can be seen as a superset of diverse sub-areas such as sensor networks, mesh networks or an enabler for pervasive computing, it has attracted significant research interests in the past several years. A major advantage of MANETs over regular wired or wireless networks is their infrastructure-less nature, as MANETs can potentially be deployed more rapidly and less expensively than infrastructure-based networks. However, the lack of an underlying explicit infrastructure also becomes a major disadvantage in adapting MANETs to a wider array of applications, since existing network algorithms and protocols are often not “plug-in” solutions for such dynamic networks. New algorithms need to be designed for fundamental network tasks such as addressing, topology discovery, and routing.

Location discovery is emerging as one of the more important tasks as it has been observed that (semi-) accurate location information can greatly improve the performance of other MANET tasks such as routing, energy conservation, or maintaining network security. For instance, algorithms such as location aided routing (LAR) (Ko, 2000), GRID (Liao, 2001), and GOAFR+ (Kuhn, 2003) rely on location information to provide more stable routes during unicast route discovery. The availability of location information is also required for geocast (multicast based on geographic information (Jiang, 2002)) algorithms such as location-based multicast (LBM) (Ko, 1999), GeoGRID (Liao, 2000) and position-based multicast (PBM) (Mauve, 2003). To minimize power consumption, the geographical adaptive fidelity (GAF) algorithm (Xu, 2001) uses location information to effectively modify the network density by turning off certain nodes at particular instances. Furthermore, in (Hu, 2003), the authors have shown that wormhole attacks can be effectively prevented when location information is available. As more algorithms are being proposed to exploit location information of devices in ubiquitous networks, it is clear that obtaining such information efficiently and accurately becomes of great importance.

A direct way of obtaining location information is to install global positioning system (GPS) receivers on devices. However, this is currently impractical as GPS receivers are still relatively expensive, power-hungry, and require clear line of sight (i.e., making indoor usage impossible) to several earth-bound satellites. In ubiquitous environments (e.g., sensor networks), devices are imagined as small as possible and operating on a very restricted power source, thus, it may not be feasible to install GPS receivers onto all sensor nodes. Localization in MANET refers to the problem of finding the locations of those non-GPS enabled nodes based on limited information such as the locations of some known beacons (also referred to as anchors) and measurements such as ranges or angles among the neighbors. In this chapter, we will study why localization is important and how it can be accomplished within the ubiquitous computing environment.

BACKGROUND

The localization problem is hard for a number of reasons:

1. **Geometric limitations**: To pinpoint its exact location in 2-D, a node needs to know the locations and its distances to at least three beacons. Alternatively, nodes could calculate their own location based on a distance and an (absolute) angle measurement from one beacon. However, even if obtaining such measurements were possible and the measurements were exact, guaranteeing that the right number of beacons surround each node is often impossible since MANETs may be randomly deployed and that in general only a small percentage of nodes are indeed beacons. Thus, localization algorithms often
need to take advantage of multi-hop information, that is, estimating node locations based on other nodes’ location estimates.

2 **Availability of measurements:** For localization algorithms that require distance or angle measurements, certain sensory devices will need to be available to provide such readings. However, it is likely that not all nodes have the same sensory capacity. In other words, there is a need for the localization algorithm to work in a heterogeneous environment, in which devices with different location sensory capacities coexist.

3 **Measurement error and error propagation:** Even when measurement devices are available, there is a general consensus that those measurements are prone to errors. For instance, a distance measurement can be derived based on a received signal strength indication (RSSI) reading, in which a receiving device measures the strength of the signal from a sending device and obtains the distance via an estimated signal propagation model. RSSI reading is prone to multi-path fading and far field scattering. The error can be especially high when there are a significant number of obstacles in between the sender and the receiver. Since most localization algorithms require measurements from nodes several hops away, the measurement error is likely to aggregate along the path and eventually completely throw off the location estimate.

Despite the difficulties listed above, there is an increasing amount research effort spent on the localization problem. The amount of effort is justified because localization is considered an enabling technology that needs to be resolved with the best possible outcome upon which other location-dependent technologies for MANETs can be successfully employed. Researchers have been working on problem in both hardware (i.e., improving measurement accuracy of devices) and software (i.e., improving the localization algorithm design). This chapter covers the latest advances in this field, including the following topics:

1. **Applications of localization:** In this section, we will establish the need for better localization techniques by surveying a number of proposed algorithms for MANETs that rely on localization.
2. **Measurement types for localization:** In this section, we will cover the latest advances in hardware design that enables localization on the smaller devices commonly seen in the ubiquitous computing environment, including the devices that measure distance ranging, angle of arrival (AoA), and interferometric ranging.
3. **Survey of localization algorithms:** We will survey some of the most popular localization algorithms, including those using connectivity information, ranging, and angle information. We will study the pros and cons of each algorithm, and suggest their appropriate applications in ubiquitous computing.
4. **Localization theory:** We will cover the theoretic basis of localization techniques. We will study the necessary and sufficient conditions for a network to be localized based on the latest results from graph theory. We will show that the localization problem in general is NP-Complete. We will also introduce the reader to the Cramer Rao Bound (CRB) that is often used to analyze the hardness of different localization scenarios.
5. **Future directions:** We will look into a number of promising future directions for the localization techniques.

**APPLICATIONS OF LOCALIZATION**

There have been numerous algorithms proposed for MANETs that rely on localization data; in this section, we provide a brief survey of these
algorithms. We will consider algorithms in four categories based on their functionalities: unicast routing; multicast routing; energy consideration; and network security.

**Unicast Routing**

Routing refers to the task of finding the correct route from a sending device (source) to a receiving device (destination). Routing is an especially challenging task for MANETs because their frequent topology change implies the underlying instability of any established routes. As such, routes are needed to be frequently rediscovered, reestablished, and repaired. In general, routing (i.e., route discovery and repair) involves flooding the routing control packets throughout the network. Flooding can often be expensive in terms of delay and bandwidth usage it incurs, both of which can greatly affect the network performance. Thus, there is a strong incentive to design efficient routing algorithms that minimize the overhead caused by any unnecessary packet flooding. Unicast routing based on location information, often called geometric routing or location based routing, has shown to be one of the viable solutions to this problem.

**Location-aided routing (LAR)** (Ko, 2000) protocol is the first MANET routing algorithm proposed that uses location data. In LAR, every node is assumed to know its own location, and each individual location is then periodically broadcast throughout the network. Thus, at any time $t$, every node knows the locations of any other nodes at some previous time $<t$. Based on this location information and an estimated velocity, a node can derive an estimated location range, called “expected zone,” of a target node at the current time. Instead of flooding the entire network, the routing request packets can be directed to search for the target node only at this expected zone. Global flooding is performed only after the location based routing request has failed. Limiting route discovery to a smaller expected zone with LAR reduces the number of routing requests compared to the standard flooding scheme.

**GRID** (Liao, 2001) protocol uses location information as a way to form geographical clusters within the network. Based on node locations and their residency within a pre-determined grid system, nodes within the same grid block are grouped into a cluster. A cluster head or “gateway” in (Liao, 2001) is then selected for each grid block. The cluster head is responsible for servicing the routing packets. Furthermore, the cluster head can monitor the status of existing routes and reroute packets as deemed necessary. Since the cluster formation effectively simplifies the network topology, the routing overhead is reduced. A critical requirement of forming such geographical-based clusters is the availability of node location information.

In (Kuhn, 2003), the authors provided a theoretical bound to the geometric routing problem and proposed an algorithm called GOAFR+. Assuming that node locations are known using some localization technique, GOAFR+ first tries to greedily route the packet by forwarding it to the neighbor located closest to the destination. However, such greedy selection does not guarantee message delivery since the intermediate node closest to the destination might not have a route to it. In such cases, GOAFR+ explores the boundaries of the faces of a planarized network graph by employing the local right hand rule (i.e., always turn right) to escape the local minimum. This method of escaping local minima is also called “parameter routing,” which is used in a number of other location based routing protocols as well.

In terms of performance, simulations performed by (Ko, 2000) and (Liao, 2001) have shown up to 50% of reduction in routing packets when using geographic routing compared to standard flooding. Since the overhead of flooding is proportional to network density, it has been observed that the amount of performance gain becomes more significant when network density
is increased. Furthermore, although the routing performance is impacted by the localization error, such impact is observed to be minimal. This indicates that in the case of routing, highly precise location data is not required. After all, location data is used by routing algorithms to give a direction that guides the routing packets; imprecise location data can still be used as long as the general direction is valid.

**Multicast Routing**

Similar to unicast routing, multicast routing can also benefit from location data. Multicast routing using geographic information is often referred to in the literature as geocast routing. The Location-Based Multicast (LBM) algorithm (Ko, 1999) is a multicast extension to the unicast Location-Aided Routing (LAR). Like LAR, which forwards the routing requests according to the location of the destination node, LBM forwards the requests according to the direction of the geocast region that contains all the multicast destinations. GeoGRID (Liao, 2000) is the multicast extension to GRID (Liao, 2001). Like in GRID, location information is used by GeoGRID to identify the grid block where nodes reside. Multicast is done through the gateway node selected at each grid block. Based on the location of the source node and the geocast region, LBM and GeoGRID define a “forwarding region” that contains the intermediate nodes responsible for forwarding packets. The size and shape of the forwarding region have a direct impact on the overall performance; shapes such as rectangles and cones have been proposed in (Ko, 1999).

While the standard shapes such as rectangles and cones work well in most cases, there are situations where viable routes exist only outside the forwarding region. For instance, a network can be partitioned into two sub-networks connected only through a narrow linkage due to some obstacles (e.g., two islands connected by a bridge). When the source and the destination are in separate partitions, a geometrically defined forwarding region is unlikely to cover the linkage. To prevent routing failure in such a case, a routing zone based on Voronoi diagrams was proposed in (Stojmenovic, 2006), which partitions the network graph based on the proximity of the nodes. Again, the proximity information relies on localization information.

The Position-Based Multicast (PBM) protocol proposed in (Mauve, 2003) attempts to optimize the multicast tree it generates by minimizing the overall path length and the overall bandwidth usage; two often contradictory objectives. To minimize the overall path length, PBM takes a greedy approach using location information. At each intermediate node, packets are forwarded to a set of neighbors based on their overall distances to the multicast destinations. In particular, a set of the neighbors with the minimum overall distance to every destination is selected as the next set of forwarding nodes. To take account of the bandwidth usage, the greedy selection also tries to minimize the size of the forwarding set. PBM also uses parameter routing to deal with local minima. Both greedy routing and parameter routing employed by PBM rely on the location information.

**Power Management**

MANET is often used as the model for sensor networks, an emerging technology for pervasive computing. One of the major challenges of sensor networks is power management. Since sensors are commonly small in size and are battery powered, conserving energy would prolong their service time and, thus, the lifespan of the entire network. The Geographical Adaptive Fidelity (GAF) algorithm (Xu, 2001) is a network topology management algorithm with reduced energy consumption as its primary objective. The idea behind GAF is that there are often a large number of nodes that are redundant during packet routing in MANETs. If the redundant nodes can be identified, they
Device Localization in Ubiquitous Computing Environments

can then turn off their radio to save energy. For GAF, the identification of redundant nodes is accomplished by analyzing the relative location information among the neighboring nodes. More specifically, GAF divides the network into virtual grids such that all nodes in grid block $A$ are the neighbors of all nodes in grid block $B$. This way, all nodes within the same virtual grid block can be considered equivalent. To conserve energy during packet routing, GAF only turns on the radio for one of the nodes in each grid block. The active node is periodically “round-robin” to achieve load-balancing. Analysis and simulations performed in (Xu, 2001) show that GAF can reduce overall energy consumption by 40% to 60%.

Security

In (Hu, 2003) the authors proposed a technique called “packet leashes” to defend against wormhole attacks in MANETs. A wormhole attack is a type of security breach where an adversary intercepts incoming packets and tunnels them to another part of the network via a single long-range directional wireless link or through a direct wired link. From there, the adversary can retransmit the packets to the network. Note that this type of “capture-and-retransmit” attack can be immune to common packet encryption methods, since the adversary does not need to read the packet content. Wormhole attacks can severely disrupt ad hoc routing protocols such as Ad hoc On-Demand Distance Vector Routing (AODV) or Dynamic Source Routing (DSR), and cause a denial of service to the network. The core of “packet leashes” is based on two assumptions: i) all nodes know their own locations; and ii) all nodes are synchronized. To enable packet leashes, the sender node encloses its location and transmission time-stamp within the packet. At the receiver node, the packet leash is validated against the receiver’s own location and clock. In particular, the sender location information gives the distance from the original sender to the receiver, and the time-stamp gives the transmission duration of the packet. Based on the transmission duration and signal propagation model, factored in some error tolerance, the receiver can validate the estimated distance the packet has traveled against the true distance to determine whether the packet is indeed coming from the original sender or an imposer at some other location. Thus, the location information and time-stamp provide a virtual leash to limit the effective range of the packet so that it cannot be exploited by wormhole attackers.

From the previous discussion on the location-dependent algorithms that encompass a wide range of problem domains, it is clear that providing location information (i.e., localization) to MANETs is becoming an increasingly important task. In fact, localization is now widely regarded as an “enabling technology” for MANETs that needs to be addressed before other location-dependent techniques can be realized in the real world (Patwari, 2003).

Measurement Types for Localization

In this section, we study a number of measurement types provided by onboard hardware devices that enable localization in MANETs for ubiquitous computing environments. A Global Positioning System (GPS) (Parkinson, 1996) receiver can provide the absolute location. However, its cost, size, and power requirement prevent it from being installed at every network node. How do nodes not equipped with GPS obtain their location information then? They have to rely on sensory measurements provided by alternative hardware devices. There are five general types of measurements: i) connectivity only; ii) RSSI (radio signal strength indicator) ranging; iii) ToA (time of arrival) ranging; iv) AoA (angle of arrival), and v) interferometric ranging. We will describe each of their capacity, usage and mathematical models when applied to the localization problem.
CONNECTIVITY ONLY MEASUREMENT

At a minimum, a node can detect connectivity to its immediate neighbors, that is, its one-hop neighborhood. The connectivity only measurement is a binary reading between two nodes of either “true” or “false” indicating whether they are neighbors. Based on this connectivity information, one can derive the general proximity of the nodes as a way to localize the network.

RSSI Ranging Measurement

A node can be localized using multilateration (Niculescu, 2001) if the distances (i.e., the ranges) to three or more known locations are obtained. The distances can be obtained, for example, by measuring RSSI (radio signal strength indicator) or ToA (time of arrival). In RSSI, the receiver measures the received signal strength and compares it with the transmitted signal strength. The difference (in dB) is then applied to the inverse of the signal propagation model to provide a distance estimate. Sensors that measure RSSI are widely available to mobile devices. Indeed, most off-the-shelf technologies implicitly provide such information (e.g., most WiFi, Bluetooth, and IEEE802.15.4 chipsets do). The drawback of RSSI-based measurements is that they can be very inaccurate because an exact model of the propagation environment is often unavailable. Experiments in (Savvides, 2001) have shown that when no obstacle exists between the sender and the receiver, RSSI can provide a distance estimate of accuracy within a few meters. However, in a less than ideal environment, the result is often unpredictable. Furthermore, low cost RSSI receivers are often variable in their transmission power due to the lack of calibration.

In the outdoor environment with a minimum of obstacles, signal propagation decay is proportional to $d^{-p}$, where $d$ is the distance the signal has traveled, and $p$ is an environment-dependent path loss exponent. However, in the actual environment where obstacles exist, multipath signals and shadowing become two major sources of noise that impact the actual RSSI. In general, those noises are commonly modeled as a random process during localization. Let $P_{i,j}$ be the RSSI (in dB) obtained at the receiver node $j$ from the sender node $i$. $P_{i,j}$ is commonly modeled as a Normal distribution (Patwari, 2003)

$$P_{i,j} = N(\bar{P}_{i,j}, \sigma_{dB}^2)$$

where $\bar{P}_{i,j}$ is the mean power in dB and $\sigma_{dB}^2$ is the variance caused by noise factor such as shadowing. $P_{i,j}$ is further defined as the power reduction from a reference location:

$$P_{i,j} = P_0 - 10n_p \log_{10}(d_{ij} / d_0)$$

where $P_0$ is the power at a reference location at the distance $d_0$ (commonly $d_0 = 1m$), $p$ is an environment-dependent path loss exponent that is assumed to be known from prior measurements (theoretically $p = 2$). $d_{ij}$ is the Euclidean distance between nodes $i$ and $j$.

ToA Ranging Measurement

Although ToA is used for radio signals in GPS, it is mostly used in the context of acoustic or ultrasonic signals in inexpensive ToA tracking (as propagation speeds are five orders of magnitude less). For instance, the Medusa node in (Savvides, 2001) is an implementation of ToA ranging using ultrasonic signals. ToA measures the time signals travel from the sender to the receiver. The distance between nodes is obtained by multiplying this time with the signal propagation speed. In spite of the additive noise and multipath, in general distance measures based on ToA are more accurate than RSSI-based measures. However, special acoustic transceivers have to be employed on each node and synchronization among the nodes needs to be established. Clock synchronization algorithms...
design for sensor networks that are accurate to the order of 10 µs have been reported (Sivrikaya, 2004). As mentioned earlier, ToA may also be used together with radio signals, but current technology is not mature enough to provide satisfactory precision over smaller distances inexpensively.

Let \( i \) be the sender node and \( j \) be the receiver node, ToA measurement \( T_{i,j} \) is often modeled as a Normal distribution (Patwari, 2003):

\[
T_{i,j} = \mathcal{N}(d_{i,j} / c, \sigma^2)
\]

where \( d_{i,j} \) is the Euclidean distance between \( i \) and \( j \), \( c \) is the signal propagation speed, and \( \sigma \) is the variance caused by noise.

**AoA Measurement**

A node can be localized if the angles between it and two beacons are known. Thus, the angle information (i.e., bearing, or angle of arrival (AoA)) can be used to localize the network. Currently, there is no off-the-self device that offers AoA sensing capability. However, a number of prototype devices are available. For instance, Cricket Compass (Priyantha, 2001) is a small form device that uses ultrasonic measurements and fixed beacons to obtain acoustic signal orientations. In (Niculescu, 2004) a rotating directional antenna is attached to an 801.11b base station; by measuring the maximum received signal strength, a median error of can be obtained from the sensor. The challenge here is to design AoA sensing devices with small form factor and low energy consumption. In (Chintalapudi, 2004), the authors outline a solution with a ring of charge-coupled devices (CCDs) to measure AoA with relatively low energy consumption.

In general, AoA is also modeled as a Normal distribution. Let the true angle between the sender \( i \) and \( j \) be \( \hat{A}_{i,j} \), the AoA measurement between \( i \) and \( j \) is therefore

\[
A_{i,j} = \mathcal{N}(\hat{A}_{i,j}, \sigma^2)
\]

where \( \sigma^2 \) is the angle variance. Theoretical results for acoustic-based AoA estimation show standard deviation \( \sigma_A \) is between 2° to 6°, depending on range (Patwari, 2005). RSSI-based AoA method with \( \sigma_A \) on the order of 3° has been reported in (Ash, 2004).

**Interferometric Ranging Measurement**

Interferometric ranging is a “widely used technique in both radio and optical astronomy to determine the precise angular position of celestial bodies as well as objects on the ground (Kusý, 2006).” Interferometric ranging exploits the property that the relative phase offset between two receivers determines their distances to two simultaneous senders. Due to the recent advancement in hardware, it is now possible to implement interferometric ranging sensors in much smaller form factor to be used for localization (Marótí, 2005). By synchronizing the transmission at the two senders, each of which sends a signal at a slightly different frequency, the receivers can derive the relative phase offset of the two signals by comparing the RSSI readings. The distance difference (also called the \( q \)-range) can then be calculated from the relative phase offset with high accuracy. A \( q \)-range obtained from interferometric ranging from two senders \( A \) and \( B \), and two receivers \( C \) and \( D \) is the distance difference

\[
d_{ABCD} = d_{AC} - d_{BD} + d_{BC} - d_{AD} + e
\]

where \( e \) is the measurement error (Figure 1). A major advantage of interferometric ranging is that the measurement could be extremely accurate compared to noise-prone RSSI readings. In a recent experiment (Marótí, 2005), in which 16 nodes are deployed in a 4x4 grid over a 18x18 meter flat grassy area with no obstruction, the maximum \( q \)-range error was shown to be around 0.1 meters while the medium error was less than 0.04 meters. However, interferometric ranging is more difficult to implement due to the following reasons.
The measurement can be impacted by various sources of noise such as frequency drift, ground multipath error, and time synchronization error (Maróti, 2005). Frequencies of the transmissions need to be precisely calibrated, as any carrier frequency drift and phase noise would directly impact the observed phase offset. Precise time synchronization is needed at the senders of a q-range. Thus, there will be overhead to maintain clock synchronization.

A significantly larger number of measurements are required for localization than using direct ranging techniques. While there are also a large number of measurements available \(O(n^4)\) even for a small network, only a small subset of them are independent of each other. The rest merely provide redundant information. It has been shown in (Kusý, 2006) that the number of independent measurement using interferometric measurements is \(O(n^2)\), which is significantly higher than with RSSI and AoA ranging \(O(n)\). Considering the localization problem in relative coordinates, for a network of \(n\) nodes there are \(2n-3\) unknowns in two dimensions and \(3n-6\) unknowns in three dimensions. This is because the relative coordinates are invariant under translation, rotation, and reflection. Thus, in two dimensions, we have \(2n-3\) degrees of freedom, where translation, rotation, and reflection each reduce one degree of freedom. Thus, the smallest network that can be localized using interferometric measurements is a fully-connected network with a population of \(n=6\), where there are 9 independent measurements available to cover 9 unknowns. The large number of q-ranges available/required indicates a scalability issue for larger networks.

Since each measurement involves four nodes, more collaboration is required between nodes. Due to the requirement of synchronized transmission, the senders have to collaborate in scheduling their transmission. Also, the receivers have to collaborate to derive the relative phase offset. Such collaboration requires sophisticated protocols to be implemented in order to reduce the communication overhead.

Those difficulties rooted in the physical characteristics of interferometric ranging devices affect the algorithmic design of the localization algorithm. As we will see in the following section, the localization algorithms based on interfero-

---

Figure 1. The interferometric ranging measurement of the q-range \(d_{ABCD} = d_{AD} - d_{BD} + d_{BC} - d_{AC} + e\). Here, node A and B are the senders, and node C and D are the receivers.
metric ranging measurements tend to be more difficult to design.

Table 1 summarizes the five measurement types described in this section.

**LOCALIZATION ALGORITHMS**

The previous section introduced the primary types of measurement that can be used for localization. However, obtaining measurements such as distance ranging and angle of arrival is only the first step of localization. To calculate the actual node location, we will have to mathematically incorporate those measurement readings to derive localization algorithms. While there are various ways of classifying localization algorithms, we feel it is more logical to classify them according to the measurement assumptions as follows: i) connectivity-only; ii) range-based; iii) angle-based; iv) interferometric ranging based; v) hybrid, and vi) mobility-based.

**Connectivity-Based Algorithms**

A number of localization methods rely on connectivity information only. These types of methods are also referred to as “range-free” methods in the literature. For instance, the Centroid method (Bulusu, 2000) estimates the location of an unknown node as the average of its neighboring beacon locations. Clearly, in order for the location estimate to be reasonably accurate, a large number of beacons need to be heard. Thus, to provide sufficient localization coverage, the Centroid method requires more powerful beacons with a large transmission range.

The APIT (Approximated Point-In-Triangulation) method (He, 2003) estimates the node location by isolating the area using various triangles formed by beacons. For each triangle formed by three beacons, the node is either in or out of the triangle. For instance in Figure 2(a), if it can be determined the node G is inside ABC and DEF, G’s location can be isolated to the shaded overlapping area of the two triangles. To determine whether a node is inside or outside the triangle, APIT

---

**Table 1. Measurement types for localization**

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Accuracy</th>
<th>Cost</th>
<th>Measured Value</th>
<th>Math Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connectivity</td>
<td>2</td>
<td>N/A</td>
<td>Proximity</td>
<td>1 – if two nodes are connected; 0 – otherwise</td>
</tr>
<tr>
<td>RSSI</td>
<td>2</td>
<td>Low</td>
<td>Distance (d_{ij}) between node (i) and (j), derived from power (P_{i,j})</td>
<td>(P_{ij} = N(\mu_{ij}, \sigma^2_{ij}))</td>
</tr>
<tr>
<td>ToA</td>
<td>2</td>
<td>High</td>
<td>Distance (d_{ij}) between node (i) and (j), derived from time of arrival (T_{ij})</td>
<td>(T_{ij} = N(d_{ij} / c, \sigma^2_{T}))</td>
</tr>
<tr>
<td>AoA</td>
<td>2</td>
<td>Low/Medium</td>
<td>Angle (A_{ij}) between node (i) and (j)</td>
<td>(A_{ij} = N(\hat{A}<em>{ij}, \sigma^2</em>{A}))</td>
</tr>
<tr>
<td>Interferometric Ranging</td>
<td>4</td>
<td>Very High</td>
<td>q-range (distance difference between four nodes)</td>
<td>q-range</td>
</tr>
</tbody>
</table>

\[ d_{ABCD} = d_{AD} - d_{BD} + d_{BC} - d_{AC} + e \]

between node \(A, B, C\) and \(D\). Noise to each q-range is modeled using a Normal distribution
compares the RSSI readings from the beacons at the node with those at its neighbors. Intuitively, smaller RSSI reading means a shorter distance (i.e., closer to the beacon) and vice versa. If there does not exist a neighbor that is further from (or closer to) all beacons simultaneously, then the node is inside triangle with high probability. For instance in Figure 2(b), a neighbor of $D$, $E$, can be measured to be further away from the beacon $A$, $B$ and $C$ because it has smaller RSSI readings comparing to $D$. Thus, $D$ is considered as to be outside $ABC$. Conversely, if $D$ is inside $ABC$ (Figure 2(c)), then it is likely that its neighbors will be closer to (or further away from) some (but not all) of the triangle points. Clearly, this test does not guarantee correctness every time. However, since there are a large number of triangles available for the test ($\mathcal{O}(n^3)$ for $n$ beacons), error can be effectively controlled. Indeed, simulations performed in (He, 2003) indicated that APIT gives more accurate localization than the Centroid method when the beacon density is higher. Note that although APIT makes use of RSSI, it is only used to derive the relative proximity, but not the absolute distance. Thus, we classify APIT as a connectivity-based algorithm.

Both the Centroid and APIT methods try to localize the node directly from the beacons 1 hop away. Thus, to provide better localization coverage, they require either a large number of beacons or an extended beacon transmission range. The DV-Hop method (Niculescu, 2001) relaxes such requirement by providing a way to localize from the beacons several hops away. In DV-Hop, each beacon floods its location to the entire network much like the distance vector (DV) routing protocol (Perkins, 1997). The algorithm contains two phases. In the first phase, a distance-per-hop estimate is obtained using DV. In the second phase, each node calculates its location estimate using the beacon locations and the distance-per-hop estimate. Each node maintains a DV table of the beacon locations it has heard along with the shortest hop count to them. A node will only forward the location broadcast if it has a shorter hop count than the current one in its table. In addition, when a beacon has heard the broadcast originated from another beacon, it can derive the distance-per-hop information based on the physical distance between the two beacons and the hop count accumulated along the path. The distance-per-hop information is then broadcast to other nodes. To localize, a node extracts the hop counts to the beacons from its DV table and converts them into distances using the average distance-per-hop information it has received. The node can then estimate its location using multilateration based on the distances to the beacons. For instance in Figure 3, the node $D$ can triangulate based on the location broadcast from the beacons $A$, $B$, and $C$ stored in its DV table. The distance-per-hop is calculated as the average of the distances per hop among all the beacons. Compared to Centroid and APIT, DV-Hop re-
requires a much lower number of beacons. It does, however, have greater communication overhead since it requires multiple message flooding.

The above connectivity-based localization methods assume the nodes are stationary. The MCL (Monte Carlo localization) method (Hu, 2004) takes a novel approach by making use of node mobility. As a node moves, it becomes connected or disconnected to other nodes. Based on the connectivity observation, a unit-disk connectivity model, and a simple random movement model of the node, MCL updates the probability distribution of the possible node location. Simulation in (Hu, 2004) has reported as many as three times of localization accuracy when compared to the Centroid method.

In general, connectivity-based localization algorithms such as Centroid, APIT and DV-Hop tend to be simple to implement, and they depend less on special hardware. However, due to the lack of more precise measurement, the location estimates they provide tend to be less accurate. A large number of beacons need to be deployed in order to improve their accuracy. However, sparse networks by nature contain less connectivity information, and, thus, they are more difficult to localize accurately using connectivity-based localization methods.

**Figure 3. DV-Hop**

$$\text{distance-per-hop} = \frac{|BC|/6 + |AB|/6 + |AC|/6}{3}$$

### RSSI and ToA Range-Based Algorithms

Many algorithms use the RSSI and ToA measurement to derive the distance to the senders. The DV-Distance method (Niculescu, 2001) behaves much like the connectivity-based DV-Hop method. But instead of incrementing the hop count, DV-Distance increments the distance between hop to hop as beacons broadcast their locations. Since the distance at each hop can be quite different, DV-Distance can obtain a more accurate range to the beacons compared to DV-Hop, which only considers the average case. However, its performance becomes dependent on the ranging measurement accuracy.

The Euclidean method (Niculescu, 2001) tries to derive the distance to a beacon that is several hops away by measuring RSSI or ToA to its neighbors. The distance is obtained by observing some simple Euclidean constraints. For instance in Figure 4, the node $D$ is several hops away from the beacon $A$. To derive its distance to $A$, $D$ obtains the distance using RSSI or ToA to two neighbors $B$ and $C$, where the distance $AB$, $AC$ and $BC$ are known. The distance $AD$ is the second diagonal of the quadrilateral $ABDC$. Depending on whether $ABDC$ is convex or concave,
two solutions of $AD$ exist. This ambiguity can be resolved by examining multiple quadrilaterals like $ABDC$. Once the distances to at least three beacons have been obtained, both DV-Distance and Euclidean method estimate the node location using multilateration.

The Collaborative Multilateration method (Savvides, 2001) is also based on multilateration from ranging. However, it allows nodes being triangulated from non-beacon nodes. Initially, all non-beacon nodes are assumed to be at some random locations. As a node receives its neighbors’ estimated locations, it tries to triangulate its new location with the least mean square error. The newly estimated location is then sent back to the neighbors for their own multilateration. The process is iterated multiple times, and the idea is that eventually the location information from the beacons will propagate to remote nodes via collaborative multilateration. However, it is foreseeable that the nodes further away from the beacons would be slow to converge. The HopTERRAIN method (Savarese, 2002) makes an improvement in this regard by using the DV-Hop method to derive an initial coarse location. It then runs the collaborative multilateration to further refine the localization results from the distance and location information from the neighbors. The n-Hop Multilateration method proposed in (Savvides, 2003) uses a bounding box model instead of DV-Hop to provide initial location estimates. For instance in Figure 5, while node $D$ is two hops away from the beacon $B$ and one hop away from the beacon $A$, it is still bounded by distance constraints. The bound on the $x$ coordinates is $[x_{D} - a, x_{D} + b + c]$, where $a = |AD|$, $b = |BC|$ and $c = |CD|$. Using this kind of geometric bounding through multiple hops, an initial location of the node can be derived.

The iterative multilateration provides a way to deal with the difficult question of how to effectively apply the beacon information several hops away. However, since it treats location estimates from non-beacons the same as beacons, the beacon information can be quickly watered down by the inaccuracy of non-beacons. The probabilistic localization method in (Huang, 2005) explicitly considers the location uncertainty of non-beacons by the means of probability distributions. In particular, each node location is not represented by a singular value, but a probability distribution in terms of particles. Initially, all non-beacons have a uniformly distributed particle distribution. To localize, nodes exchange their particle distributions among the neighbors and run Monte-Carlo filtering based on the RSSI or ToA measurement data to update the particles. Eventually, the particles will be refined to the true location of where the node resides. The particle filtering method allows collaborative localization as shown...
in Figure 6. Here, nodes 2, 3, and 4 are beacon nodes, while nodes 0 and 1 are non-beacons. Of the beacons, node 0 can receive signals only from nodes 1 and 4, and node 1 can receive signals from only nodes 0, 2, and 3. From the signal strength readings, non-beacons estimate their distances to their neighbors. The probability distribution of the estimated location is represented by the particles (dots) in the graph. In sub-figure (a), where node 1 is removed, node 0 can only receive signals from node 4; thus, as the particle distribution indicates, the probability distribution where node 0 is most likely located concentrates on a circle around node 4. In sub-figure (b), where node 0 is removed, node 1 can receive signals from nodes 2 and 3; thus the most likely locations for node 1 center around two areas where “transmission circles” around node 2 and 3 intersect. Intuitively, in order to localize itself, a node needs to receive location information from a minimum of three beacons either directly or indirectly. In both case (a) and case (b), the exact location of the nodes 0 and 1 cannot be deduced because they do not receive location information from all three beacons. In (c) and (d), where all nodes are available, nodes 0 and 1 are able to communicate to each other and exchange their particle distributions. Thus, their probability densities will represent their actual locations much closer even though neither node receives location information from all three beacons directly.

This section introduced range-based localization methods. Compared to range-free methods, range-based methods give more accurate location estimates when ranging data is reliable. However, depending on the deployment environment, ranging techniques based on RSSI tend to be error-prone and strong filtering is required. The ranging error could ultimately throw off the localization accuracy if it is allowed to propagate through the network unbounded. Furthermore, different methods generally exploit the trade-off between the estimation accuracy and the estimation coverage. For instance, given the same network scenario, the Euclidean method is capable of generating more accurate location estimates of a smaller subset of nodes, whereas the DV-Hop method has better coverage but worse accuracy. Regardless of the tradeoff, a common characteristic shared by many range-based localization algorithms is that they require a relatively high network density in order to achieve better results. Based on the extensive simulation of DV-Distance, Euclidean and multilateration methods performed in (Chintalapudi, 2004), it can be concluded that those range-based localization algorithms “require an average degree...
of 11-12 nodes within the ranging neighborhood in order to achieve 90% localization coverage with 5% accuracy (Chintalapudi, 2004).”

** AoA-Based Algorithms **

Even though the future of AoA sensing devices is still unclear, some works have been published on localization using angle information. Simulation studies in (Chintalapudi, 2004) also show that when AoA of the signals is used in addition to the distance measurement, the localization accuracy and coverage can be drastically improved. This should not come as a surprising conclusion, as nodes need to communicate with only one neighbor to perform localization if they can obtain
both AoA and distance measurements. The work in (Chintalapudi, 2004) also presents three variations of a weighted mean square error algorithm that localizes the nodes, each of which is designed to work with one of the three measurement types: i) distance-only measure; ii) distance plus a more accurate AoA measure (up to of precision); and iii) distance plus a less accurate AoA measure (up to of precision). The less accurate AoA measurement method is sometimes referred to as *sectoring*. Simulations in (Chintalapudi, 2004) show that the localization accuracy and coverage can be greatly improved even with such coarse sectoring measurement as well.

In order to localize with only AoA measurement, the AoA triangulation method proposed in (Niculescu, 2003) can be used. The triangulation takes several AoA measurements from beacons and estimates the node location with least square error. To propagate the AoA measurement for more than one hop, the AoA triangulation method uses a method called *orientation forwarding* that is similar to the Euclidean method for distance ranging. For instance in Figure 7, let AoA measurement be the bearing against South. For node $D$ to derive its bearing to the beacon $A$ (i.e., $D$), it can contact two neighbors, $B$ and $C$, with known AoA measurements from the beacon $A$ (i.e., $B$ and $C$ are known). Furthermore, $B$, $C$, and $D$ can measure the AoA of each other to give the readings of $B$, $C$, $D$ and $D$. From there, all angles in $\Delta ABC$ and $\Delta BCD$ can be determined. The bearing from $A$ to $D$ can be derived as $DD$, where $D$ is known, and $\angle CDA$ can be determined from $\Delta ABC$ and $\Delta BCD$. Using the orientation forwarding method, the bearing to beacons can be propagated through multilops, which can then be used to triangulate from remote beacons. However, much like the case of distance propagation, measurement error becomes aggregated at each hop. Simulations in (Niculescu, 2003) have reported a near linear error increase to the hop count.

In summary, due to the limited availability of AoA sensing devices, relatively few algorithms have been proposed for AoA. However, it is conceivable that some localization algorithms originally proposed for RSSI or ToA ranging can be adapted to AoA. For instance, the probabilistic algorithm in (Huang, 2005) can be updated to accept AoA measurements by simply providing an alternative measurement model for AoA during particle filtering.

**INTERFEROMETRIC-RANGING BASED ALGORITHMS**

Due to the fact that interferometric sensing devices for localization are relatively new, there have been only a limited number of localization

---

*Figure 7. Multihop Distance Derivation in AoA Triangulation*
algorithms proposed for this type of measurement. When compared to RSSI/ToA ranging and AoA, all of which involve two nodes for each measurement, interferometric ranging involves four nodes for each measurement and, thus, makes it more difficult to propagate location information through multihops. To eliminate the multihop propagation issue, a simple genetic optimization approach was taken in (Maróti, 2005), which propagates all interferometric readings within the network to a centralized location and runs a genetic algorithm to find the node locations that match the readings. Such an approach is of more theoretical (and prototyping) interest than practical use, since any centralized method is not scalable to large networks. A Pair-wise Distance method was proposed in (Patwari, 2006) that uses both interferometric and RSSI ranging. The method uses the interferometric ranging to derive pair-wise distances among the nodes. The node locations can then be optimized using the least square error method from the pair-wise distances. The algorithm then repetitively applies the RSSI ranging measurements to fine-tune the location estimates. Compared to the genetic algorithm, the Pair-wise Distance method is able to converge much faster. However, it is currently still a centralized algorithm, which presents the same scalability issue as the genetic algorithm.

Both of the above algorithms try to optimize for a global solution given an entire set of interferometric measurements. Intuitively, finding a global solution to the localization problem is often difficult because of the large search space and the large number of constraints given by the interferometric measurements. Thus, it is desirable to find solutions in some subspaces first and then incrementally build up to the global solution. For instance, an iterative approach has been proposed in (Huang, 2007) that localizes from a small set of seeding beacons. At each round, a set of nodes that can hear from the seeding beacons are localized. As additional nodes are localized at each round, they act as pseudo-beacons that allow other nodes to be localized at subsequent rounds. The iterative method is fully distributed. However, error propagation can be an issue since any localization error at pseudo-beacons would adversely affect the localization result at subsequent rounds. Simulation results in (Huang, 2007) have shown a linear increase of localization error at each round.

**Hybrid Algorithms**

A combination of the above techniques can be employed to form hybrid localization methods. For instance, a hybrid method is proposed in (Ahmed, 2005) that uses both DV-Distance (Niculescu, 2001) and Multi-Dimensional Scaling (MDS) (Shang, 2003). The algorithm contains three phases. In the first phase, a small subset of nodes is selected as reference nodes. In the subsequent phase, the reference nodes are then localized in relative coordinates using MDS. The final phase uses DV-Distance to localize the rest of the nodes in absolute coordinates. The rational behind such hybrid algorithms is to exploit the tradeoff between different localization algorithms. For example, MDS gives good localization accuracy, but as the network size is increased, MDS can be costly. Meanwhile, DV-Distance is less costly, but it only works well when beacon ratio is high. With the hybrid algorithm, the cost is minimized by only running MDS on the reference nodes, and then the reference nodes are used as beacons for DV-Distance.

**Localization Using Mobility**

While most previous methods assume stationary beacon locations, an alternative method is to localize devices using a mobile beacon. In this method, a mobile beacon travels through the deployment area while broadcasting its location along the way. Devices localize themselves by monitoring information coming from the beacon. A straightforward technique using the above method is de-
Device Localization in Ubiquitous Computing Environments

scribed in (Ssu, 2005), where devices are required to receive at least three communications with the same RSSI reading from the beacon. Given that the same RSSI reading implies similar distances to the beacon locations, the physical device location can be derived using simple geometric functions. This method is computationally simple, making it suitable for resource-limited sensors. However, it requires the beacon to directly pass by the ranging area of the device. In addition, in most cases, the beacon has to pass by the device twice because the sampling positions of the beacon when the three RSSI readings are taken should not be on the same line. This method also assumes that errors are insignificant in the RSSI to distance translation.

Instead of computing the location directly, a probabilistic approach may be taken; here device location is viewed as a probability distribution over the deployment area. In (Sichitiu, 2004), devices measure a series of RSSI readings from the mobile beacons and localize themselves by a sequential update process to the probability distributions of their locations. Each device starts with a uniform distribution covering the entire deployment area. As the beacon passes through, the distribution is updated to fit the received RSSI readings (using a signal propagation model). The method is further improved in (Peng, 2005) by adding the negative information (i.e., the information that the beacon is out of range), as well as RSSI readings from the neighbors. These probabilistic methods provide with much improved location estimates, but have the drawback of being complex. For a deployment grid of \( n \) by \( n \) units, the time and space complexity is \( O(n^2) \). As the devices such as sensors at present time have very limited resources, it is difficult to directly implement these methods for large sensor deployment scenarios. Indeed, the experimental results shown in (Sichitiu, 2004) are performed on pocket PCs, which are much more powerful than cheap devices like sensors.

A similar method of localizing the networks using a mobile beacon is presented in (Galstyan, 2004). Instead of the actual probability distribution, the possible device locations are represented with a bounding box. As the beacon passes by, the area contained by the bounding box is progressively reduced as positive and negative information is processed. The bounding box method drastically simplifies the probability computation, making it possible to implement this method on sensor devices. However, such large simplification has its side-effects in that it sacrifices the preciseness of the distribution for its simplicity as the box cannot precisely describe multiple possible locations. There is also the problem of noise from ranging devices. This method may work well when ranging error is minimal; however, when noise is present (which is inevitable when using RSSI ranging), there might be situations where no bounding box exists to satisfy all readings.

Table 2 lists all the localization algorithms described in this section. In summary, different measurement types and their unique properties to a large degree dictate the design of localization algorithms. For instance, connectivity-based measurements can only provide coarse localization without a higher beacon ratio or nodal degrees. Range and AoA-based measurements can provide much finer localization results, but they are more prone to measurement error. A quantitative comparison between the more well-known algorithms such as DV-Hop, Euclidean and multilateration can be obtained from (Langendoen, 2003), in which the comparison is done in the context of specific constraints of sensor networks, such as error tolerance and energy efficiency. Their results indicate that there is no single algorithm that performs “best” and that there is room for further improvement.

THEORETICAL RESULTS

While there have been many localization algorithms proposed for various scenarios, only recently have researchers started to address the
Device Localization in Ubiquitous Computing Environments

Table 2. List of localization algorithms

<table>
<thead>
<tr>
<th>Connectivity-based</th>
<th>Centroid (Bulusu, 2000), APIT (He, 2003), DV-Hop (Niculescu, 2001), MCL (Hu, 2004)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range-based</td>
<td>DV-Distance (Niculescu, 2001), Euclidean (Niculescu, 2001), Collaborative Multilateration (Savvides, 2001), Hop-TERRAIN (Savarese, 2002), n-Hop Multilateration (Savvides, 2003), Probabilistic Localization (Huang, 2005)</td>
</tr>
<tr>
<td>Angle-based</td>
<td>Weighted Mean Square Error (Chintalapudi, 2004), AoA Triangulation (Niculescu, 2003)</td>
</tr>
<tr>
<td>Interferometric Ranging</td>
<td>Genetic Algorithm (Maróti, 2005), Pair-wise Distance (Patwari, 2006), Iterative (Huang, 2007)</td>
</tr>
<tr>
<td>Hybrid</td>
<td>DV-Distance + Multi-Dimensional Scaling (MDS) (Ahmed, 2005)</td>
</tr>
<tr>
<td>Mobility-based</td>
<td>Geometric Localization using Three RSSI Readings (Ssu, 2005), Sequential Update to Probability Distribution (Sichitiu, 2004).</td>
</tr>
</tbody>
</table>

Theoretical aspects of the localization problem. In this section, we briefly cover the latest theoretical results with regard to the localization problem. Since ubiquitous computing environment is often modeled as a graph, it is not surprising that much of the theoretical work is based on graph theory. With regard to localization, we are particularly interested in the following three theoretic problems: i) localizability; ii) complexity of localization; and iii) localization error bounds.

First of all, we would like to know that given a network scenario (i.e., the nodes and their relative measurements such as ranging and angling) whether it is theoretically possible to uniquely localize the network. Such knowledge of localizability is important to us because if we can easily identify the scenario that is impossible to localize uniquely, then it would be pointless to run any localization algorithm on it. Instead, we would have to request additional nodes or measurement data to be available (by possibly deploying more nodes or beacons) so that the localizability requirement is satisfied. The following theorem gives the necessary and sufficient condition for distance-constrained network localizability in two dimensions.

**Theorem 4.1** The network is localizable in two dimensions if and only if the network graph is redundantly rigid and triconnected (Hendrickson, 1992; Berg, 2003).

The above theorem makes use of some graph theory concepts. In graph theory, rigidity (or first-order rigidity) in general refers to the situation in a graph where there are no continuous motions of the vertices satisfying distance constraints on edges. A graph is redundantly rigid (or second-order rigid) if the induced graph remains rigid after removing any single edge.

Theorem 4.1 holds for two dimensions only. The sufficient condition for higher dimension is currently unknown. To test the localizability, there exists a polynomial time algorithm (O(n²) where n is the number of nodes) that tests for the first-order rigidity; see (Hendrickson, 2002) for one implementation. However, it is a known NP-Complete problem to test for the second-order rigidity of a graph (Saxe, 1979). A related but even more difficult problem is node localizability, which asks if a particular node (instead of the entire network) is localizable. No sufficient condition of node localizability is currently known even in the two dimensional case and, thus, no deterministic algorithm currently exists.

A second problem asks for the theoretic complexity of localization itself. In particular, we would like to know that given a network scenario that satisfies localizability whether there exists a deterministic polynomial time algorithm that would localize the network. This problem deals with the NP-Completeness of localization. Unfortunately, the hardness of graph realization
has been shown as NP-Complete under the measurement of distance (Eren, 2004), angle (Bruck, 2005), connectivity (Breu, 1998; Kuhn, 2004), and interferometric ranging (Huang, 2007).

The above theoretical results indicate the general intractability of the localization problem even in the ideal case where measurements (such as edge distances) are 100% accurate. Unfortunately, measurements in the real world are a far-cry from being accurate, and any optimization method has to deal with not only different measurement types, but also noise. The localization inaccuracy attributed to the measurement types and noise can be statistically qualified using Cramer-Rao Bounds (CRB) (Patwari, 2005). The CRB is a lower bound on the covariance of any unbiased location estimator that uses measurements such as RSSI, ToA, or AoA. Thus, the CRB indicates a lower bound of the estimation accuracy of a given network scenario regardless of the localization algorithm. In other words, with CRB we have a way to tell the best any localization algorithm can do given a particular network, measurement type, and measurement noise scenario. CRB formulas of individual measurement types such as RSSI, ToA, and AoA under most common noise models (mostly Gaussian) are currently known.

The CRB of the localization error for a sample network is shown in Figure 8 as rings of radius being the standard deviation of the minimum localization error that can be possibly attained at the node. Here, the nodes represented by squares are beacons while circles represent nodes to be localized using RSSI ranging. The edges indicate the communication links available to measure RSSI readings. We assume the measurement model to be RSSI with the path loss exponent $p$ and the standard deviation of the noise $dB$. A ring with smaller radius (i.e., a smaller CRB) signals that more accurate localization result can be theoretically obtained. Conversely, a larger ring indicates a larger localization variance and, thus, a less accurate result. In the figure, two types of nodes do not have rings. First, all beacons have a CRB of 0. There are also regular nodes that have infinite CRB indicating that those nodes are theoretically impossible to localize. The latter case can be seen at nodes 38, 48, 49, and 78 in the top left corner. At a minimum, three beacons are needed to localize a connected network. However, those nodes in the top left corner are isolated to a different partition. Since they are connected to only one beacon (node 91), those nodes clearly cannot be localized. Other than those cases, the CRB rings at the main network partition clearly show the level of localization difficulty under various scenarios. In general, we observe that nodes closer to the beacons tend to have a smaller CRB than the ones that are several hops away. Even smaller CRB can be obtained when a node is closer to more than one beacon. All of the above observations are consistent with our common intuition about localization difficulty.

It is important to note that CRB is essentially a theoretic bound that depends on the measurement model. In the real world, its usefulness is limited by how accurate the measurement model reflects the reality. Nevertheless, CRB can be a useful tool in comparing various localization algorithms. It can be used to validate how close a particular algorithm can come to this theoretic lower bound and to see if there is any room for improvement in the algorithm design.

**CONCLUSION**

In this chapter, we studied the localization problem in ubiquitous computing environments. Localization in general refers to the problem of identifying the physical location of devices using a limited amount of available measurement data. The most common measurement types include device connectivity (i.e., whether two devices are neighbors), ranging using RSSI and ToA, angle of arrival (AoA), and interferometric ranging. Given a small number of nodes with accurate geometric location (e.g., using GPS receivers), localization
algorithms try to derive the location of those devices that are not “GPS-enabled.” The motivation of localization can be justified by the large number of algorithms proposed for ubiquitous computing that rely on (semi-)accurate location information and the fact that current technology prevents GPS from being installed on all network devices due to power constraints and form factors. It has been shown that localization in general, regardless of the measurement types used, is an NP-Hard problem. Thus, current effort in solving it relies on some sort of stochastic optimization. Meanwhile, as with other network-related problems in ubiquitous computing environments, the ideal solution calls for a distributed but efficient implementation, which leads to additional challenges.

Like other aspects of ubiquitous computing, the localization problem is relatively new. The problem is also conceptually straightforward to characterize, and many results from other disciplines such as graph theory, optimization theory, online algorithms can be readily applied to this problem. Thus, researchers from other disciplines can provide valuable insight that could lead to better solutions to the problem. It is our hope that this brief introduction will provide the readers motivation and inspiration to perform research in this exciting field.

Figure 8. The CRB of the sample network is depicted as rings of the radius i. There are two exceptions: 1) beacons, depicted as squares, have 0 CRB, and 2) some regular nodes have infinite CRB (such as node 38, 48, 49 and 78 at the top left corner) indicating that they cannot be localized.
Device Localization in Ubiquitous Computing Environments

FUTURE DIRECTIONS

Device localization within ubiquitous computing environment has been an active research field in the past several years. Much work has been done in the area of hardware/sensor design (in particular, reducing the form factor and power consumption of sensory devices), algorithmic design and theoretical analysis. However, like many areas of ubiquitous computing, localization is still a relatively new front with much of the work yet to be done. In this section, we will briefly discuss a few directions which we feel could produce fruitful results in the near future. We hope our discussion will encourage the readers to actively participate and contribute their own ideas to this exciting and important field.

Implementation and Testing Environment

When reviewing the previous works on localization, one cannot help but notice a disturbing trend; a majority of works on localization have been based on either theoretical models or numerical simulations, while works based on the result of actual hardware implementation have been relatively few. It is not difficult to project that the primary reason for such trends is the hardware cost. To perform meaningful experiments for localization, especially for those collaborative localization methods such as DV-Distance and Euclidean, one would normally need a large number (100+) of devices. Although simple mobile ad hoc network devices (e.g., sensor motes) are becoming cheaper by the day, it is still quite costly to implement algorithms on physical devices on such a large scale. In addition, the sensing capacities of the current devices are usually limited to RSSI. Currently, there is no cheap hardware that implements AoA, ToA, or interferometric ranging and, thus, most works using these measurement types are all based on simulations. In a sense, the advances in algorithmic work on the localization problem are currently outpacing the advances in hardware. Future work needs to be done to significantly improve the hardware design to fill this gap.

Another issue related to the testing environment is that there is no common localization test bed. While large scale network simulators such as NS2 have modules for simulating mobile ad hoc networks, these modules do not contain localization. Smaller simulators for sensor networks such as SENSE (Chen, 2004) usually lack features on localization as well. Another simulator SENS (Sundresh, 2004) does explicitly implement localization, but it uses a rather primitive Centroid method and lacks more sophisticated methods. Since NS2 is the most widely-used network simulator, it would be very helpful for researchers to implement an interface to NS2 that allows a “plug-in” for future localization algorithms. This would give a common test bed for different localization schemes. More importantly, it would also allow those location-depended algorithms (such as location-aided routing methods) to be implemented and compared based on the result of localization. The localization module should also implement the Cramer Rao Bounds (CRB) so that the theoretic error bound can be calculated for different localization scenarios.

Interferometric Ranging

Since interferometric ranging is a relatively new type of measurement available to the localization problem, there are still many open problems in this area. Of the localization algorithms proposed for interferometric ranging, all but the iterative algorithm proposed in (Huang, 2007) is centralized. There is a definite need to design distributed localization algorithms for interferometric ranging so that it can be implemented with reasonable efficiency and scalability. To reduce the number of beacons, the distributed algorithms should make use of multi-hop location information, which unfortunately is much more difficult for interferometric ranging because each measure-
Device Localization in Ubiquitous Computing Environments

The localization process involves four nodes. For instance, there is a scheduling issue, both at a high level and low level, of when a device should be scheduled to send or receive interferometric readings. At a high level, since not all q-ranges are independent, it is more desirable to schedule the senders and receivers in order to generate more independent q-ranges. At the low level, after the senders and receivers are selected, they have to be scheduled to coordinate the signal transmission because the transmission needs to be synchronized. The design of the scheduling algorithm can have a substantial impact on the overall performance of the localization algorithm in terms of the localization accuracy and communication overhead.

Furthermore, a simulation study in (Huang, 2007) has shown that the multi-hop error propagation has a big impact on interferometric ranging, which increases almost linearly as the localization results are propagated at each hop. Therefore, the control of the error propagation is another research issue. There is also a need for an algorithmic independent theoretic error bound (like CRB) for interferometric ranging. The bound would be more difficult to derive than those for distance ranging and angling because more than two nodes are involved in each measurement. Thus, the inter-dependence between the error and the relative locations of senders and receivers becomes more challenging to characterize mathematically. However, the payoff of obtaining such bound is that it would allow us to ultimately compare interferometric ranging with other measurement types and identify the scenarios that are preferable for each measurement type.

Collaborative Localization of Multiple Measurement Types

Previous localization algorithms often assume that the entire network has to be localized using the same type of measurement (such as connectivity-only, RSSI, ToA, AoA, or interferometric ranging). However, to be true to the spirit of ubiquitous computing, it is foreseeable that future networks will consist of devices of vastly different capacities in terms of i) different transmission coverage, ii) power requirement, and iii) measurement sensors. Thus, during localization it is often desirable to explicitly consider various devices capacities in such heterogeneous networks. For instance, in terms of power requirements, the localization algorithm should exploit the devices with more power capacity and try to minimize the calculation performed on less powerful devices. Furthermore, different measurement types have different error characteristics. It would be interesting to investigate how to collaborate multiple measurement types during localization, and in particular how the collaboration would impact the localization error. Intuitively, incorporating multiple measurement types in the same localization scenario should improve performance since such collaboration can potentially cover the drawbacks of individual measurement types for each other. Unfortunately, such collaboration could mean an increased complexity of the localization algorithms since heterogeneous networks would invalidate some assumptions often made by simple localization algorithms (such as uniform transmission range). While it is worthwhile to consider collaborative localization algorithms, it is equally imperative to keep the localization overhead under control.

REFERENCES


Device Localization in Ubiquitous Computing Environments


### ADDITIONAL READING


Device Localization in Ubiquitous Computing Environments

Wireless Communications and Mobile Computing, 6(2), 247-258.

Chapter 3.12
Socio-Cultural Interpretations to the Diffusion and Use of Broadband Services in a Korean Digital Society

Dal Yong Jin
Simon Fraser University, Canada

ABSTRACT
This chapter attempts to ascertain the causes of the rapid growth of broadband services in the context of the broader socio-cultural elements. It recognizes technology as a socio-cultural product which has historically been constituted by certain forms of knowledge and social practice, so this chapter explores cultural elements contributing to the diffusion of broadband services in the context of the cultural environment in Korea. Further, this chapter discusses the significant role of the people, as users, in the process of the rapid diffusion and growth of broadband services. In particular, it emphasizes the way in which the 1997 economic crisis, as one of the most significant socio-cultural turning points in modern Korean history, has influenced the deployment of broadband services as high-speed Internet connections have developed since 1997.

INTRODUCTION
The widespread availability of broadband services (high-speed Internet) throughout the world significantly influences people in their life. Broadband services have made it easier to download digital music and movies, and the wide penetration of broadband services has enabled customers to engage in online stock transactions and online games. Many schools, from elementary to universities, in various countries are also connected to broadband services and utilize information technology (IT) for education, while Internet broadcasters have rapidly become popular.

A wide range of developed countries, such as the United States and the United Kingdom to developing countries in the Third World, have initiated modern telecommunications networks that support broadband access, providing high-speed access and always-on connections to both homes
and offices, to develop the social and technical landscape of cyberspace (Han, 2003; Ishii, 2003; Lee & Lee, 2003; Jin, 2007). Korea is among the leading performers in broadband (Lee, Oh, & Shim, 2005). As of January 2006, about 77% of Korean households were connected to broadband services, which is one of the highest throughout the world (MIC, 2006).

Consequently, Korea has become the world’s best laboratory for broadband services—and a place to look to for answers on how the Internet business may evolve (Taylor, 2006). Much scholarly analysis and discourse (Yun, Lee, & Lim, 2002; Lee, O’Keefe, & Yun, 2003; Reynolds & Sacks, 2003; Choudrie & Lee, 2004; Lau, Kim, & Atkin, 2005; Lee et al., 2005; Jin, 2005) has focused on the roles of the government and competition among telecommunications companies in developing broadband services in Korea. They emphasize several factors contributing to the rapid growth of broadband services caused by the deregulation and competition policies in the telecommunications sector, such as a variety of promotion policies to boost Internet use and the strategies of broadband Internet providers. Previous studies, however, have not paid much attention to socio-cultural factors, which would be one of the most important contributing factors given that people are the major users of broadband services. Although a few papers examine cultural factors that contribute to broadband services, their discussions are neither comprehensive, nor informative.1

Unlike these preceding studies, this chapter attempts to ascertain the causes of the rapid growth of broadband services in the context of the broader socio-cultural elements. It recognizes technology as a socio-cultural product which has historically been constituted by certain forms of knowledge and social practice, so this chapter explores cultural elements contributing to the diffusion of broadband services in the context of the cultural environment in Korea. Further, it discusses the significant role of the people, as users, in the process of the rapid diffusion and growth of broadband services. In particular, it emphasizes the way in which the 1997 economic crisis, as one of the most significant socio-cultural turning points in modern Korean history, has influenced the deployment of broadband services, as high-speed Internet connections have developed since 1997.

**TECHNOLOGY AS CULTURAL FORMS**

It is generally recognized that technologies are primarily neutral because they operate essentially under the same norm of efficiency in all situations. Many users of technology argue that technology is essentially amoral and an entity devoid of values (Rescher, 1969; Mesthene, 1970). This instrumental theory, the dominant view of modern governments and the policy sciences on which they depend, argues that “if people use technology for destruction or pollution, as in the case of nuclear weapons and chemical pollution, it should not be blamed on technology, but on its misuse by politicians, the military, big business and others” (Pacey, 1983, p. 2).

For many scholars, however, technology is not simply a means to an end, but has become an environment and a way of life; this is its substantive impact (Borgmann, 1984). This substantive theory of technology holds that technology is not neutral, but has a substantive value bias. Substantive theory, best known through the writings of Jacques Ellul, Arnold Pacey, and Martin Heidegger, claims that technology constitutes a new type of cultural system that restructures the entire social world as an object of control. Substantive theory explicates cultural aspects of technology, such as values, ideas, and the creative activity of technology (Feenberg, 1991). This type of cultural system is characterized by an expansive dynamic which ultimately mediates every pre-technological enclave and shapes the whole of social life.
Among these, Ellul (1964) pointed out that we have to examine technology through its sociological aspects: “We should consider the effect of technique on social relationships, political structures, and economic phenomena.” Technology is not aloof from the social realm; on the contrary, it is an integral part of the social (Webster, 2002). Pacey (1983, pp. 4-6) also discussed how those who write about the social relations and social control of technology should be made aware of the cultural aspects as well as the organizational and technological aspects.

In agreement with Pacey, Christians (1989, pp. 124-125) observes: “Technology is the distinct cultural activity in which human beings form and transform natural reality for practical ends with the aid of tools and procedures.” He clearly argues that cultures are humans’ distinctive and immediate environment built from the material order by men and women’s creative effort. Furthermore, Williams (1992, pp. 127-129) stated: “How technology develops is not only a matter of some autonomous process directed by remote engineers, but is a matter of social and cultural processes.” Technologies are developed and used within a particular social, as well as economic and political context (Franklin, 1999). As these theoreticians emphasized, technology does not develop independently, but is part of a particular social-economic and cultural process. This chapter aims to contribute to this ongoing debate of substantive theoretical discourse with a case study of the rapidly growing broadband services in Korea.

**Historical Background Information on the Growth in Broadband Services**

The rapid deployment of broadband services in Korea began in 1995 when the government enacted the Framework Act on Information, which established the first Master Plan, which set up a comprehensive strategy for the Korean Information Infrastructure (KII) (MIC, 2004). The goal of the KII was to construct an advanced nationwide information infrastructure consisting of communications networks, Internet services, application software, computers, and information products and services (Jin, 2007). The KII project aimed at building high-speed networks by providing over 80% of households with more than 20Mbps (megabits per second) broadband access by 2005 through market competition and private sector investment (Lee et al., 2003).

In the post-1997 economic crisis, efforts bringing about the development of broadband services have accelerated. The economic crisis severely affected the Korean economy due to unprecedented rates of unemployment, corporate bankruptcies, and the demise of the stock market. Korea’s economic condition demanded that Korea change its industrial structure from traditional heavy and chemical industries to a more IT-oriented structure—telecommunications and computers. More importantly, Korea identified information technology as its path to economic recovery, following the 1997 financial crisis (Kim, 2006a). The Korean government has exerted great effort on deploying a high-capacity backbone and pursued a policy of high-speed telecom infrastructure as a foundation for the benefits of IT changes for sustainable economic growth (Lee et al., 2003).

Under these circumstances, both government planning and stable funding had played significant roles in the rollout of broadband services. The Korean government invested a total of $11 billion into broadband services between 1998 and 2002, while the U.S. government planned to invest only $2 billion in the form of tax breaks as of May 2003 (Belson & Richtel, 2003, p. C1). The Korean government has also initiated a variety of promotion policies expediting the growth of broadband services, such as creating Internet-Friendly Classrooms and Ten Million People Internet Education. Clearly, Korea’s rapid and expansive broadband growth can be traced, in large measure, to government initiatives (Lau et al., 2005, p. 351).
Competition among telecom companies was also dedicated to the deployment of broadband access throughout Korea. Broadband services in Korea were first introduced in 1998, and a few telecom companies, including Korea Telecom (KT), the largest telecom company in Korea, Thrunet, and Hanaro, had severely competed with each other in the market. In May 1999, Thrunet, the introducer of broadband services in Korea, held 63% of the market share, followed by Hanaro (35%) and KT (2%). The market has undergone substantive change in recent years. For example, Hanaro Telecom, controlled by U.S. investors American International Group Inc. and Newbridge, absorbed Thrunet to be in a better economic position to compete with the dominant market player KT in late 2005. LG Powercomm, a subsidy company of the LG group, entered the market in September 2005, and cable system operators have expanded their share in the high-speed Internet market (Kim, 2006b). As a result, as of April 2006, KT as the largest player consisted of 49.9% of the market share, followed by Hanaro (28.5%), cable system operators (10.6%), Powercomm (4.4%), and others (Park, 2006, p. 16).

The intense competition among telecom companies led to price (monthly fee) reductions and subsequently a rapid increase in demand for broadband Internet services. In other words, competition among telecom companies for increasing their respective market share has resulted in a drop in the price of broadband subscriptions, and it has certainly contributed to the rapid spread of broadband in Korea. Fierce competition resulted in prices decreasing from $40 per month in April 1999 to $30 in February 2003, and again to less than $20 in May 2006 (ITU, 2003a; Taylor, 2006).

Meanwhile, as a new market trend, Korea has been developing the next generation in high-speed Internet services. The country acknowledges the existence of market saturation, so it has enhanced the capacity of existing broadband services, including its speed and convergence with various types of media. Korea has begun to upgrade its high-speed Internet from less than 5 Mbps to 100 Mbps since 2005 in order to enjoy high-quality Internet telephone service, movies, and online education (Bang, 2005). Further, Korea expects the rapid penetration of 100 Mbps broadband services will expedite the growth of image-related contents, and thereafter the second wave of high-speed Internet services in 10 years. As such, the Korean government along with telecom companies played a pivotal role in the growth of broadband services by providing infrastructure and government initiatives.

**Cultural Characteristics Contributing to the Deployment of Broadband Services**

Although the Korean government and telecom firms have collaboratively initiated and expedited the deployment of high-speed Internet services, the swift growth of broadband services would have been impossible without people readily accepting new technology more than other nations. The diffusion of services and the widespread use of high-speed Internet can be attributable to certain distinct characteristics of the Korean people, since the diffusion of broadband services took place with rapid acceptance by most Koreans (Jin, 2007). Regardless of the fact that many countries around the world have initiated and supported the growth of broadband services, the result varies markedly due to the acceptance of new technology by people, the users, as well as government initiatives.

Therefore, it is crucial to explore the significance of assessing the culture in which technology is created and the context in which it is widely accepted. As ITU acknowledges (2003b), Internet user demand indeed contributed most decisively to the rapid explosion of broadband in many places, particularly in Korea. Again Ellul (1964) emphasized that “the development of technology is not an isolated fact in society but is related to every factor in the life of modern humanity.” In particu-
lar, cultural forms rooted in Korean history after
the 1997 economic crisis should be emphasized
as significant elements which prompted the rapid
deployment of broadband, because members of
the public are the major customers of broadband
services. There are several significant cultural
characteristics (i.e., the mentality of the people)
rooted in Korean society and its historical context
that are important contributing factors to the de-
velopment of broadband, such as rapid acceptance
of new services and technologies, enthusiasm for
'edutainment,' keeping up with their neighbors
(“me-too” culture), as well as developing dual
personalities of two competing forces, social
solidarity and individualism.

The primary characteristic among most Kore-
ans is demanding quick change, which eventually
expedited the growth in high-speed Internet. Only
a few decades ago Koreans were characterized
as calm and patient, as one distinguished Con-
fucian characteristic, but in just a few years, the
demand for quick change became one of the most
distinctive Korean characteristics (Han, 2003, p.
19). Since the 1960s when the country began its
high-speed transition from an impoverished state
to one of Asia’s major economies, Koreans cannot
wait for long periods of time, nor are they very
patient. Koreans have not been known for taking
things slowly. Instead, Korean society is well
known for its impulsiveness—that is, desiring
quick communication, quick contact, and quick
results (Jin, 2007). Most Koreans would certainly
agree that Korea is nothing if not dynamic. In this
regard, Jong Suk Lee (2000), director of the Korea
Culture and Policy Institute, states: “Nine out of
ten Koreans rush into the road as soon as the light
changes to green and they cross the street on or
even before the green light is about to turn red. In
contrast, Chinese would usually wait until all the
traffic is cleared and safe even at the clear green
light. And if the light is about to turn red, they
just stop and wait for the next green light.”

Indeed, “balli, balli”—Korean for “hurry,
hurry”—is heard incessantly on the crowded
streets of Seoul, reflecting the frenetic pace of
Korean life in a global era (Jin, 2007).

Another survey conducted in the 1980s is still
compelling to explain the same trend. Accord-
ing to the survey (Chinese Cultural Connection,
1987), regarding Confucian dynamism which
measures long-term vs. short-term orientation,
Japan showed one of the highest scores, compared
to the U.S., the UK, and Australia. The high Japa-
nese score in the survey reflected the patience or
long-term orientation of Japan compared with
other developing countries (McFadyen, Hoskins,
& Finn, 1998). However, Korea has lost patience
and has become addicted to short-term results,
and this changing Confucian dynamism in Korea
has contributed to the rapid penetration of high-
speed Internet.

This relentless drive, which has led Korea to
chalk up a number of significant achievements,
has driven the rapid spread of broadband Internet
connections. Most Koreans have hurried to have
broadband Internet connections, and it has been
easier to develop these services compared to other
countries, including Japan. What most Koreans
wanted was the same thing: quick communication,
quick games, and quick contact. Everything needs
to go “balli, balli,” and this unique characteristic
of Koreans has expedited the rapid penetration of
high-speed Internet (Shim, 2006). In other words,
for most Koreans, everything needs to be done
right now. For example, in Korea, instant mes-
sage-swapping is comparatively popular mainly
because not answering a message from a friend
with all due speed is considered a faux-pas in
Korean society (Taylor, 2006).

This quick pace has particularly accelerated
with the 1997 economic crisis because Koreans
began to realize that quick change and adoption
of new technology and information was needed
to survive in the globalization era. Korea is full
of ‘early adopters’ who are willing to buy newly
released digital devices for consumer testing. In
a survey by Korean ad agencies in 2004, 43% of
respondents said that they consider themselves
‘early adopters’ in the market (Kwon, 2006). According to a survey by the U.S. market research firm Parks Association in November 2005, Korea ranked second on the list of 13 countries in the adoption of consumer technologies (Parks Associates, 2005). Although balli balli culture has several negative aspects, such as bribery scandals and the collapse of buildings and bridges, in the midst of achieving social and economic successes in the Korean society, rapid adaptability to change is the key to the swift deployment of broadband services.

An excessive enthusiasm for edutainment—the combination of education and entertainment—has also greatly contributed to the unique growth in broadband services. Above all, Korea is one of the most developed countries in terms of education. Its overall school enrollment rate (primary, secondary, and tertiary) of 90% is the highest throughout the world. Korea’s high rate of literacy and school enrollment are essential as prerequisites for the widespread adoption of ICTs, and these factors have helped contribute to the growing impact of ICT in Korean society (ITU, 2003a). Over-enthusiasm for education in Korea is not new, but after the 1997 economic crisis, parents started to realize the potential threats and opportunities of globalization. It became a must for their children to acquire both English and Internet skills in order to survive in the era of globalization (Aizu, 2002). Many parents who have children younger than early teens, for example, have focused on cyber-based English education since the late 1990s. They have acknowledged that their children are able to enjoy online reading, music, game, and animation written in English through the Internet due to high speed and their being 24/7 support services, which is a good alternative to learning English in a classroom. In the Korean context, excessive need for education has also developed so-called edutainment, as a form of studying English while enjoying playing in cyberspace (An, 2001).

Edutainment is not only for children. Many mid-careers have also turned their eyes on the Internet in the midst of a changing working environment. In Korea, the majority of companies have begun to adopt a Western-style five-day workweek system since the late 1990s and early 21st century. The spreading five-day workweek system since then has boosted information technology, such as DVDs, mobile games, and edutainment. With more spare time available, people began to look for a new type of education containing more entertainment factors, instead of the existing cramming type of learning (Cho, 2002). Therefore, having high-speed access to the Internet at home became an advantage for everyone from children to housewives to career persons, or a disadvantage if they did not have it (Jin, 2007). Broadband is an indispensable component for edutainment, which is one of the major contributing factors in broadband services.

Meanwhile, the peculiar Korean “me-too culture,” which means all companies and people end up with the same systems and taste, has substantially contributed to the spread of broadband access. Consumers in various societies are susceptible to social pressure to keep up with their neighbors, (e.g., “keeping with the Jones’s” in the
U.S.); however, Koreans’ eagerness to keep up with their friends and neighbors is comparatively very high (Jin, 2007). As discussed, Koreans tend to be early adopters of technology; however more they significantly are fast followers, as Stephen Ward, consultant manager of the telecommunications group at Deloitte observed: Koreans are always conscious of the need not to be left behind by others, and the young have a great desire to conform with the gadget-carrying norm of their peers (Ward, 2004). The majority of Koreans have not admitted any difference in several sectors. Businessmen should get the same salary if they begin the job in the same company at the same time, and parents cannot accept the fact that their children are left behind others. They must be in the best position in terms of the quality and quantity of their houses, education, and income. Korean people buy new refrigerators when their neighbors buy them, and they buy pianos when their neighbors buy them, although they have refrigerators and pianos good enough to use.

Likewise, many Koreans are eager to have high-speed Internet because they want to keep up with their neighbors. They feel that if they have the service, they could be part of the group; if not, they could not be part of the group. Technology and fashion fads tend to spread like wildfire in Korea. If one individual adopts an innovation, then others feel a social pressure to follow (Bajaj, 2002; Lau et al., 2005). The me-too culture, which is despicable in some senses, works in broadband services. Many old and new apartments are installing broadband Internet connections because the construction companies and apartment complexes do not want to be behind other apartment buildings (Cho Han, 2006). People do not want to be left behind in broadband services because they do not like a feeling of uneasiness as in many other areas. The disdainful me-too culture has ironically contributed to the rapid deployment in broadband services.

Finally, a growing double personality between social solidarity and individualism, in particular in the younger generation, has played a crucial role in the rapid deployment of high-speed Internet. As well known, Korea was a country which was proud of its tradition of social solidarity (Crotty & Lee, 2002). People do not want to be left out. Due to this, the New York Times reported in 2002, “Korea is a group-oriented society, where socializing in bunches is the preferred form of interaction, and Western-style individualism is frowned upon,” while reporting booming online games (French, 2002, p. 8). To take one documented case, after conducting a case study of online games in Korea, Chee (2006) also emphasized the importance of social solidarity of teens and twenties in the process of growing online game communities. Indeed, it is very plausible because the flourish of the online communication matched the culture and emotion of the Korean people and their need to communicate with one another through boards, which are forums to discuss and exchange information, and clubs, which brought together people with similar hobbies (Seong, 2006).

However, this conventional wisdom has changed in recent years, and individualism—the degree to which individuals are able to achieve identity and status on their own rather than through membership in groups (McFadyen et al., 1998)—is on the rise. In this regards, the New York Times did not reflect the rapidly changing environment in Korea. With the swift change in the economy before and after the 1997 financial crisis as well as recent modernization, Korea’s social and cultural situation has shifted on a large scale. The younger generation increasingly seeks Westernized culture, regarding an individual’s success, private property, and nuclear family as important, instead of the value of the Confucian culture emphasizing patriarchy and large families (Jin, 2007). According to a marketing report by SK Telecom, “Those in their 20s and 30s, the first Internet generation in Korea, gives priority to their own time and space, rejecting all invasions of privacy. They are also changing the culture of collectivism to one of individualism” (Joongang
The growing individualism has resulted in a unique culture for developing broadband services and demanding personal time, space, and privacy.

The double personality of many Koreans, on one hand social solidarity and on the other hand growing individualism, has ironically contributed to the rapid growth of broadband services. As Chee (2006) observes, young people enjoyed popular online games sometimes with friends and other times with nobody, but through high-speed Internet. This shows social solidarity on the one hand; however, it also explains growing individualistic mentality, which prefers enjoying games with no virtual friends in cyberspace to playing with one’s peers on the playground. While seeking social solidarity, many young people enjoy high-speed cyberspace by creating their own world, and this changing social cultural characteristic has become crucial for the rapid growth of high-speed Internet.

In sum, the explosion of broadband in Korea has been possible due in large part to various cultural factors, rooted deeply in Korean society and its historical context, in particular the 1997 economic crisis. Korean citizens are the most significant component in broadband diffusion because they are the main users of the service. Without high-speed Internet, nobody can have a normal socio-cultural life, including enjoyable individual cyberspace from online games to edutainment, in contemporary Korea, which is one of the fastest growing information societies. As Aizu (2002) points out, the social and cultural factors, the aggressive mentality of the Korean people, high awareness of the challenges of globalization, and political and historical contexts played the decisive role in its dynamic acceptance of the Internet and acceleration to broadband in Korea.

CONCLUSION

Korea presents a unique example with its rapid deployment of broadband penetration. Several significant factors have contributed to the rapid development of broadband Internet connections. The government and telecommunications companies, as providers, have played important roles in the rapid development of broadband Internet, in particular to providing infrastructure for its development. Favorable government policies and competition among telecommunications companies became driving forces for the rapid deployment of broadband Internet.

The explosion of broadband in Korea, however, was greatly made possible against the backdrop of the 1997 economic crisis and due in large part to various deeply rooted social, historical, and cultural factors. Although political factors can be important driving forces behind broadband penetration, growth also requires an existing receptiveness to using the services and applications that can be provided through broadband (Jin, 2007). In this regards, it was the citizens who actually made inroads into the world’s most shrewd market for broadband services in Korea.

Several socio-cultural factors, which are crucial to the diffusion and use of new technologies, have played significant roles in the swift deployment of broadband services. Cultural characteristics emphasizing quick communication and quick responses, as well as enthusiasm for edutainment, have contributed to the exponential growth of broadband services. A growing complexity of characteristics of the younger generation showing double personality of social solidarity and individualism, which is not easy to find in any other country, has particularly contributed to the rapid growth in broadband services. If the younger generation has only one characteristic of these
two, it could not be the strong cultural factor for the rapid growth of broadband services. In the midst of globalization that expedited with the economic crisis in 1997, the younger generation has developed its unique characteristics, which resulted in one of the most significant cultural factors in broadband Korea. The majority of Koreans have been eager to gain new technology, and their unique cultural characteristics have become the crucial initiative of the rapid deployment of broadband services.

In conclusion, the introduction and development of new technology cannot be explained solely in terms of political structures and economic policy surrounding technology. As Ellul (1964) stated: “One should be looking at technology in its sociological aspect because technology is not an isolated fact in society but is related to every factor in the life of modern man.” Technology does not develop independently, but is part of a particular social-economic and cultural setup. The deployment of broadband in Korea has been possible due to its specific socio-cultural dimensions as well as political and economic environment. Cultural characteristics rooted in the Korean society greatly influence the growth of technology and broadband services throughout the country.

REFERENCES


Socio-Cultural Interpretations to the Diffusion and Use of Broadband Services in a Korean Digital Society


KEY TERMS

1997 Economic Crisis: The Korean economy has gone from being an example of one of the most successful development experiences in modern history to economic stagnation and decline in 1997. From the middle of 1997, Korea was beset with a series of financial crises. The trend of decades of rising incomes reversed, and unemployment and poverty were reaching alarming levels. In particular, factors responsible for the decline in the value of exports then include a dramatic fall in the prices of some electronic and information equipment, in particular semiconductors, which had dire consequences for a number of countries in the East and South Asia region.

Confucianism: A philosophy of life developed by Confucius. It stressed the proper relationships in society, such as father/son and ruler/subject. The philosophies of Confucius are emphasizing love for humanity; high value given to learning, and devotion to family, peace, and justice.

Convergence: Also known as digital convergence. The concept that all modern information technologies are becoming digital in nature. The technological trend whereby a variety of different digital devices such as TVs, in particular high-definition TVs, and mobile telephones are merging into a multi-use communications appliance employing common software to communicate through the Internet.

Edutainment: The combination of education and entertainment. Many people in the digital age use the Internet as a form of studying several subjects, in particular English while enjoying playing in cyberspace.

Information Technology (IT): Compared to the labor-led technologies and industries, information technology usually includes semiconductors, computers, and telecommunications, although some economists like to use the term “knowledge-based industry” in order to explain IT.
Online Games: Internet games (also known as online games) are games that are played online via the Internet. They are distinct from video and computer games in that they are normally platform independent, relying solely on client-side technologies. Normally all that is required to play Internet games are a Web browser and the appropriate “plug-in.”

Social Solidarity: The degree or type of integration of a society. In simpler societies, solidarity is usually based on kinship ties or familial networks. In more complex societies, solidarity is more organic, referring to the interdependence of the component parts. Thus, social solidarity is maintained in more complex societies through the interdependence of its component parts.

ENDNOTES

1 For example, Yun et al. (2002) and Lee et al (2003) mention that the strong emphasis on education and academic performance has prompted parents to turn to the Internet for educational goods and services; however, they do not extend their discussion further.

2 Since broadband penetration was above 77% as of January 2006, Korea almost achieved one of the major goals; however, the speed of broadband, another major goal, could not be fulfilled given that an average speed of broadband was only 2.5Mpbs in May 2005 (see Kim, 2005).

3 In 1997, Korea experienced the worst economic recess in modern history, mainly due to the crisis in financial and corporation sectors. Thirteen large conglomerates, including Daewoo and Donga, collapsed or became partially insolvent because of the shortage of currency. The immediate result of these companies’ collapse was the suffering of small companies and workers. When they failed, thousands of small-scale companies collapsed. As of February 1998, 3,377 mid-size and small companies were bankrupt. When companies went bankrupt, workers lost their jobs. By the end of June 1999, the number of jobless reached 1.78 million. This means the unemployment rate rose sharply from 2.6% in the fourth quarter of 1997 to 8.5% in the second quarter of 1999, and real wages fell by 7.9% (see Jin, 2007).

4 The government placed an emphasis on creating Internet-Friendly Classrooms at every level of schooling by connecting them to high-speed Internet free of charge as the main component of educational information infrastructure. The government spent $1.4 billion for the establishment of school information infrastructure between 1998 and late 2000. Of this amount, the government used $600 million to connect classrooms of 10,064 elementary, middle, and high schools around the country with broadband services (see Lim, 2001, p. 3).

5 The relatively low monthly subscription fees have helped the growth of broadband services; however, it is almost certain that a high subscription fee could not deter the spread of broadband services because the majority of Koreans have not taken the price as a factor. For example, most Koreans over teens have had cell phones regardless of high monthly subscription fees. The subscription fees of mobile phones have greatly risen since 1997; however, the penetration rate has also exponentially increased. In the late 1990s, the Korean government officially banned mobile carriers from offering handset subsidies in an attempt to block excessive marketing; however, many local mobile users still wanted to have mobile phones, although the full price of many handsets were priced at more than $500 in 2002. There are major factors other than promotion policies and
competition among telecom companies in the Korean context which would be cultural factors.

6 100Mbps broadband services can download the equivalent of a 32-page newspaper per second and a film (1.4 GB) within two minutes, compared to one hour with 3.5 Mbps broadband services. About 80% of customers used broadband services of 1-4Mbps in Korea as of December 2005.

7 Parks Associates surveyed more than 10,000 households in 13 countries and ranked nations according to their proclivity to adopt and use MP3 players, video-on-demand (VOD), home networks, computers, online services, and similar advanced technologies. The survey found that Taiwan (Digital Living Index: 8.7) and Korea (8.4) lead the world in the adoption of consumer technologies, followed by the U.S. (8.1), Canada (8.0), and Japan (7.9) (see Parks Associates, 2005).

8 It is in some ways despicable because Korean people follow others, not because they need, but because they want to be in the same class and enjoy showing off.

This work was previously published in Handbook of Research on Global Diffusion of Broadband Data Transmission, edited by Y. Dwivedi; A. Papazafeiropoulou; J. Choudrie, pp. 78-89, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 3.13
Mobility and Multimodal User Interfaces

Christopher J. Pavlovski
IBM Corporation, Australia

Stella Mitchell
IBM T. J. Watson Research, USA

INTRODUCTION

Traditional user interface design generally deals with the problem of enhancing the usability of a particular mode of user interaction, and a large body of literature exists concerning the design and implementation of graphical user interfaces. When considering the additional constraints that smaller mobile devices introduce, such as mobile phones and PDAs, an intuitive and heuristic user interface design is more difficult to achieve.

Multimodal user interfaces employ several modes of interaction; this may include text, speech, visual gesture recognition, and haptics. To date, systems that employ speech and text for application interaction appear to be the mainstream multimodal solutions. There is some work on the design of multimodal user interfaces for general mobility accommodating laptops or desktop computers (Sinha & Landay, 2002). However, advances in multimodal technology to accommodate the needs of smaller mobile devices, such as mobile phones and portable digital assistants, are still emerging.

Mobile phones are now commonly equipped with the mechanics for visual browsing of Internet applications, although their small screens and cumbersome text input methods pose usability challenges. The use of a voice interface together with a graphical interface is a natural solution to several challenges that mobile devices present. Such interfaces enable the user to exploit the strengths of each mode in order to make it easier to enter and access data on small devices. Furthermore, the flexibility offered by multiple modes for one application allows users to adapt their interactions based on preference and on environmental setting. For instance, hands-free speech operation may be conducted while driving, whereas graphical interactions can be
adopted in noisy surroundings or when private data entry, such as a password, is required in a public environment.

In this article we discuss multimodal technologies that address the technical and usability constraints of the mobile phone or PDA. These environments pose several additional challenges over general mobility solutions. This includes computational strength of the device, bandwidth constraints, and screen size restrictions. We outline the requirements of mobile multimodal solutions involving cellular phones. Drawing upon several trial deployments, we summarize the key designs points from both a technology and usability standpoint, and identify the outstanding problems in these designs. We also outline several future trends in how this technology is being deployed in various application scenarios, ranging from simple voice-activated search engines through to comprehensive mobile office applications.

BACKGROUND

Multimodal interaction is defined as the ability to interact with an application using multiple sensory channels (i.e., tactile, auditory, visual, etc.). For example, a user could provide input by speaking, typing on a keypad, or handwriting, and receive the subsequent response in the form of an audio prompt and/or a visual display. Useful multimodal applications can cover a broad spectrum including tightly synchronized, loosely synchronized, and complementary modes of operation. Synchronization behavior must be defined both for input (the way in which input from separate modes is combined) and for output (the way in which input from one mode is reflected in the output modes). The W3C distinguishes several types of multimodal synchronization for input as follows (W3C, 2003a):

- **Sequential**: Two or more input modalities are available, but only a single modality is available at any given time.
- **Simultaneous**: Allows input from more than one modality at the same time, but each input is acted upon separately in isolation from the others.
- **Composite**: Provides for the integration of input from different modes into one single request.

A general framework for multimodal systems is depicted in Figure 1. This diagram elaborates further on several fundamentals positioned by W3C.

The interaction manager is responsible for combining multiple requests, dialog management, and synchronization. The function of receiving and combining multiple inbound requests is the responsibility of the integration manager sub-component. Conversely, the generation manager is responsible for distributing multimodal output to all of the respective output channels (modes) via an interpretation layer, which may involve text to speech (TTS) conversion or transcoding of graphical content to accommodate the needs of the target modality. Earlier work in multimodal systems referred to the integration tasks relating to composition and decomposition of requests as fusion and fission respectively (Coutaz, Nigay, & Salber, 1993).

Speech-based telephone interfaces currently available in the commercial market commonly use varying levels of directed dialog. Directed dialog, as the name implies, employs a style of system prompts that helps to “direct” the user in what to say next. Users are often presented with spoken menu options from which they can make a selection, thus navigating in a controlled manner until the task is completed. Much of the naturalness and power of speech is undermined when the application relies too heavily on the use of directed dialogs. A Natural Language speech interface, which allows the user to phrase their
request in a wide variety of ways, reduces the
cognitive load since there are no commands to
memorize or hierarchies to navigate. A mixed-
initiative interface allows the user to share con-
trol over the direction of the dialog, making the
interaction more efficient for the user.

Device manufacturers can install special-
ized software or firmware on handsets to enable
distributed speech recognition (DSR). DSR tech-
nology digitizes the speech signal and sends it
over an error-protected data channel to a speech
recognizer on a server. Thus the processing is
distributed between a terminal client and a server.
The accuracy of speech recognition can be bet-
ter when using DSR because the degradations
associated with sending speech over the mobile
network, such as low bit rate speech coding and
channel transmission errors, are avoided. In ad-
dition, DSR allows for the implementation of
a multimodal, speech and data, application on
devices which do not support simultaneous voice
and data connections.

MULTIMODAL TECHNOLOGY
IN MOBILITY SYSTEMS

Multimodal applications for small mobile devices
must overcome several technical challenges;
these include device capability and computational
strength, functional and bandwidth constraints of
the network, and limitations of the user interface.
Present work in multimodality is focused upon the
support of two modes of interaction, most typically
data (graphics or text) and speech (Kondratova,
2004; Pavlovski, Lai, & Mitchell, 2004a; Hastie,
Johnston, & Ehlen, 2002; Kvale, Warakagoda, &
Knudsen, 2003). Future trends support additional
modes that include visual gesture, lip reading,
and haptic responses. In this section we present
advances in multimodality supporting speech
and data entry, outlining the current state of the
technology used and the outstanding challenges
yet to be addressed.

Multimodal Architectures

Due to the need to support two or more modes of
input and output, the solutions to support multi-
modal systems are more complex than unimodal
systems. Additional capabilities are required to
support composite input and output requests, 
manage multiple application states, and perform
session management between devices and the
multimodal application services.

There are fundamentally two architectural
approaches to constructing multimodal solutions
for mobile devices such as mobile phones and
PDAs. The most widely investigated architecture appears to involve deployment of an application onto the mobile device, using distributed speech recognition (DSR) and a GUI client application. Such a solution may be categorized as *distributed client architecture* (i.e., thick client architecture). The alternative employs the *browser-based architecture* (thin client), where a standard Web browser is the only client software used.

Several proposed solutions have been studied that support distributed client architectures (Kvale, Narada, & Warakagoda, 2005; Klante, Krösche, & Boll, 2004). Figure 2 provides an overview of the key sub-systems and components of the distributed client architecture. The device is required to host several components of the solution, generally including a multimodal client application and distributed multimodal processing technologies. This may include automated speech recognition (ASR), text to speech (TTS), and handwriting recognition (HR); however, typical distributed client designs to date have such components on the server. Communication between the client and server occurs over a single channel; hence the multimodal input is marshaled (encoded) into one communications stream.

There are several advantages of this type of architecture including true synchronicity in delivery of the multimodal output, distributed processing, and richer user interface. Since the multimodal output is combined before transmission to the client, the client application is able to ensure media synchronization—that is, synchronized presentation of content to the user; this is a key problem to address in multimodal systems. The presence of a client application also provides a richer graphical user interface and consistency in user interface output. Additionally, the capability to distribute some processing tasks to the client provides greater flexibility for multimodal applications, particularly in relation to response time. For instance, a multimodal short message service is able to perform speech recognition locally within the device, rather than returning to the server, and display the utterance to the user before sending as a short text message.

A key drawback to distributed client architectures is the need to manage the client software to be deployed to the mobile device. This is appears to be of particular concern to mobile operators for several reasons. Given the significant numbers of mobile phones in use, software distribution and management to the device is costly. The majority of mobile phone plans are 12 or 18 months, hence new devices are acquired rapidly and continual support becomes cost prohibitive. Additionally, it may not be practical to support all devices available on the market.

Further research suggests that while a limited graphical user interface is of concern to users, in fact the major contributor to improved usability of multimodal user interfaces is the speech inter-
face (Hastie et al., 2002; Pavlovski et al., 2004a). There is also much work on the notion of speech-centric multimodal user interfaces (Johnsen & Kvale, 2005). These observations also support the notion that alternative architectures require consideration.

An alternative approach is a class of multimodal system that does not require client applications to be deployed to the mobile device. Such a framework is also termed a browser-based, or thin client, architecture—only requiring a standard browser on the mobile phone. In order to support the multiple channels, typically speech and data, use is made of a ‘Class A’ mobile phone and the multi-call supplementary service, which is only available in 3G networks. Specifically, each device establishes several bearer channels with the server, one per mode of user interaction. For example, simultaneously a voice circuit is established for speech and a data bearer service for the browsing graphical and text content (see Figure 3).

Thin client architectures have been studied for mobile devices. An initial set of building blocks to support browser architectures for sequential multimodal processing was outlined in Niklfeld, Finan, and Pucher (2001). The approach in Chou, Shan, and Li (2003) and Li, Wong, and Guo (2004) supports a thin client architecture, where use is made of the multi-call supplementary service for simultaneous connections between the mobile phone and the multimodal application server; however, only sequential multimodal input and output is supported. A further solution solves the problem of supporting composite multimodal output response (Pavlovski, Wood, Mitchell, & Jones, 2004b), however only sequential simultaneous multimodal input is supported. The MONA project provides a similar capability for thin client, and extends the device range to support both mobile phones and PDA devices using a multimodal presentation server (Anegg, Dangl, & Jank, 2004).

The browser-based architecture would naturally seem more attractive to mobile phone operators given the dynamic and rapid turnover of mobile phones, hence eliminating the need to manage software distribution and complexity attributed to supporting numerous mobile devices. The disadvantage of course is that coordinating additional channels increases the complexity and introduces the possibly of latency or synchronization errors. One paper points out that media synchronization is an impact to multimodal applications (Pavlovski et al., 2004b). This may be accommodated by delaying one of the responses so that both arrive simultaneously at the client device. However, complex multimodal applications are yet to be studied to comprehensively assess the impact of synchronization.

Several key challenges remain for browser-based multimodal architectures. This includes management of channel latency and media syn-
chronicity. A full implementation that supports both input and output multimodal composite requests remains as an outstanding problem. And, support for three or more multimodal channels is yet to be explored. These problems may be solved by using the distributed client architecture and have been demonstrated (Johnsen & Kvale, 2005)—however, at the cost of introducing additional software management overheads and restrictions on device supportability.

**Standards and Technology**

The W3C organization has defined several multimodal standards, including interaction requirements and an interaction framework (W3C, 2003b). We briefly summarize some further technologies that are specific to multimodal solutions, outlining their applicability to mobile devices.

**Extensible Multimodal Annotation (EMMA)**

EMMA is a set of XML specifications that represents multimodal user input including speech, pen, and keystroke. The XML is intended to be generated automatically, hence is suitable for use between recognition engines and the interaction manager.

**Multimodal Presentation Markup Language (MPML)**

The MPML Language is designed for presenting multimodal output on browsers supporting XML. The limited processing capability of phones is accommodated, with a mobile edition defined for J2ME applications (Saeyor, Mukherjee, Uchiyama, & Ishizuka, 2003). As such, the architecture resembles a distributed client solution.

**Speech Application Language Tags (SALT)**

SALT is a set of extensions to HTML (or XHTML and WML) adding speech capability to these mark-up languages. While it is suggested to support the thin client architecture, applicability to mobile phones is limited due to the requirement of browsers’ support for SALT. The need to manage multiple application versions, one for each device, is also observed as a drawback (Kondratova, 2004).

**XHTML+Voice (X+V)**

X+V is a proposed standard for multimodal markup. Designed for clients that support spoken and visual interaction, X+V technology furnishes traditional Web pages with further voice tags for input and output speech tasks. A goal is to support thin client, however in its current form this technology is most suited to the distributed client architecture. Use of VoiceXML has also been studied (Niklfeld et al., 2001).

**Synchronized Multimedia Integration Language (SMIL)**

SMIL is a relevant standard to managing media synchronization, an important multimodal problem. However, the standard is largely aimed at conventional browser technology that supports XML, DOM, and XHTML. Hence, it is restricted to browsers that may not be deployed to mobile phones.

**User Interface Design**

There is some work treating multimodal user interface design in mobility. A framework for
multimodal interface design has been proposed in Baillie, Simon, Schatz, Wegscheider, & Anegg, 2005). The authors suggest the need to study user behavior in the natural environment. This follows the observation that user behavior differs considerably when comparing a controlled environment to the general freedom of use that mobile devices provide (Baillie & Schatz, 2005). A key feature of the method is a table designating the context in which the multimodal system is used and the users’ preferred mode of interaction. Further work extends these concepts by proposing a tool to more fully analyze user behavior over an extended period of time (Salembier, Kahn, Calvet, Zouinar, & Relieu, 2005).

Given that there is little data on user behavior in a mobile environment, early design efforts will require careful consideration with users on how to maximize multimodal capabilities. Several key factors determine user behavior including response to error recovery, the course of action context, and properties of the implementation of each modality (Calvet, Julien Kahn, Pascal Salembier, & Zouinar, 2003). Further work shows that the speech interface appears to be the major contributor, when complementing a text or graphical interface, to the improvements bestowed through multimodality (Pavlovski et al., 2004b).

Summarizing the literature, the key features that a multimodal user interface design must address include the capability to dynamically alter modes, support for simultaneous composite output, and the ability to review the output request on one mode, using an alternative mode prior to actioning the request; for example, users who may dictate a message would prefer to see a transcript of the message prior to sending.

**FUTURE TRENDS**

There are several industry trends and future areas of research including the extension of multimodal solutions to accommodate haptic response, visual gesture recognition, lip reading, and speech technologies that support natural language understanding. Oviatt (2003) reviews several of these emerging technologies.

The idea of haptic response is already in use with present-day mobile phones, where mobile phones offer a silent mode that alerts the user with a vibration event. Such stimuli has been in use for some time and appear with early flight control systems that provide feedback to the pilots’ control system indicating turbulent flight conditions. More recently, gaming controllers use such feedback. There is some recent work on the design of haptic response (Kaaresoja & Linjama, 2005) in mobile devices, while other work focuses on haptic input to control a mobile device (Oakley & O’Modhrain, 2005). The notion of extending applications in mobility with such haptic responses is relatively new. To extend such a capability to multimodal applications, access to native mobile services such as vibration is required to initiate a haptic response.

Lisetti et al. (2003) describe research progress on a multimodal emotion recognition system. They have developed a prototype that takes input from multiple modes (i.e., camera, microphone, wearable computer) and includes both subjective, as expressed by the user, and physiological components as inputs to its model. Lip reading technology, as an alternative interaction mode, may be used to improve speech recognition. Prior research suggests that bimodal audiovisual speech recognition can improve as much as 40% in accuracy (Chan, 2001), in comparison with using audio input alone. Additional work on the use of Natural Language understanding to augment the multimodal user experiences has been proposed and studied (Pavlovski et al., 2004b).

A final observation is required regarding security and privacy. Traditionally, voice communications are considered relatively secure, due to the transient nature of circuit establishment between parties and the real-time exchange of data between two peer entities. This is no longer the case for
multimodal application, where the utterances are digitally recorded and stored. Furthermore, in order to improve the speech accuracy and fault resolution, there is a desire on the operators’ part to store such data for a perceived user benefit. Of course the storage of personal conversations and requests for information raises several security and privacy issues that have yet to be addressed.

CONCLUSION

The literature has shown that multimodal user interfaces are able to provide a superior user experience when interacting with multimodal applications on mobile devices such as cellular phones. Furthermore, the applicability of multimodality is extending beyond traditional mobile Web applications to a greater range of domains. For instance, due to the plurality of user interface modes, multimodal systems provide a natural improvement for people with disabilities (Kvale et al., 2005). Although studies demonstrate that multimodal user interfaces provide a superior user experience on a mobile phone, it remains unclear whether such an improvement remains consistent for all types of applications.

Distributed client architectures are able to overcome several problems identified in the literature, however the practicality of deployment for mobile phones requires consideration. Browser-based multimodal architectures overcome deployment issues and provide an alternative and easier mechanism for allowing mobile devices to access services. In order to support multimodal applications that make use of three or more interaction modes, the use of thin client architecture is as yet unproven. In addition, further capabilities that may not be inherent within the devices are required, such as local composite management. Hence, distributed clients would seem a necessary choice to support these capabilities and appear most appropriate to address the needs of the increasing complexity due to several modes of interaction, particularly as the media synchronization of modes becomes more complex.

REFERENCES


Mobility and Multimodal User Interfaces


**KEY TERMS**

**Automated Speech Recognition (ASR):** The use of computer processing to automatically translate spoken words into a text string.

**Directed Dialog:** Speech interface where the user receives system prompts that direct the user on available options that usually require a response with simple spoken words.

**Fission:** The decomposition of output from a multimodal application, to be distributed to the different modalities.

**Fusion:** The composition of input from multiple modalities into one input request to the multimodal application.

**Multi-Call Supplementary Service:** Enables several simultaneous circuit-switched calls to the same device, each call using its own dedicated bearer. Only one circuit-switched bearer can be used for speech at any one time.

**Multimodal Interaction (MMI):** W3C term for a class of human to computer interaction that involves multiple modes of user interaction, such as speech, writing, text, and haptics.

**Text To Speech (TTS):** The use of computer processing to transform text into spoken words.

*This work was previously published in Encyclopedia of Mobile Computing and Commerce, edited by D. Taniar, pp. 644-650, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 3.14
Foreseeing the Future Lifestyle with Digital Music: A Comparative Study Between Mobile Phone Ring Tones and Hard-Disk Music Players Like iPod

Masataka Yoshikawa
Hakuhodo Inc., Japan

ABSTRACT
This chapter aims to explore the future trajectory of enjoying digital music entertainment among consumers comparing the characteristics of the usage patterns of digital music appliances in the U.S. and those in Japan. As the first step of this research, the author conducted two empirical surveys in the U.S. and Japan, and found some basic differences in the usage patterns of a variety of digital music appliances. Next, a series of ethnographical research based on focus-group interviews with Japanese young women was done and some interesting reasons of the differences were discovered. In Japan, sharing the experiences of listening to the latest hit songs with friends by playing them with mobile phones that have the high quality, ring tone functions can be a new way of enjoying music contents, while hard-disk music players like iPod have become a de facto standard of the digital music appliances in the world.

INTRODUCTION: CENTRAL QUESTIONS
The November 2001 debut of iPod and the subsequent opening of iTunes Music Store have brought a rapid expansion of the digital music market around the world. Some estimate that the market will be worth $1.7 billion dollars by 2009 (Jupiter Research). Now, iTunes Music Store service is available in 30 countries around the world, with the total number of downloaded songs surpassing the 500 million mark in July 2005.

The store only opened in Japan in August 2005 and sold over 1 million songs in the first 4 days. This is an astonishing achievement, consider-
Foreseeing the Future Lifestyle with Digital Music

ing that Japan’s largest online music store Mora has monthly sales of around 450,000 songs. In March and April 2005, SONY, which has long led the portable music player market, released a new digital music player under the Walkman brand, offering both the hard disk type and USB flash memory type to launch a marketing drive against iPod. The developments have finally begun to provide Japanese music lovers with an environment whereby digital music contents are broadly enjoyed in terms of both services and hardware devices.

One of the major characteristics of Japan’s digital music market has been the presence of digital music contents for use on mobile phones. The use of digital music contents on mobile phones, which started as regular ring tones, has gradually evolved into Chaku-uta® (true-tone ring tones) by December 2002, and to Chaku-uta Full™ (mobile-phone-based music distribution service launched in December 2004 by the mobile carrier “au”). Chaku-uta® and Chaku-uta Full™ have sold over 100 million songs and 10 million songs respectively, making the digital music service the largest segment in mobile-phone content services.

The environment for enjoying digital music content is set to expand even further into the future. How would such a development affect the way Japanese music fans listen to music in general? This paper examines future ways of enjoying digital music content in Japan, and the competition between music players like iPod for use with personal computers and mobile phones that have adopted the usage as music players.

JAPAN’S DIGITAL MUSIC CONTENT MARKET AND THE PROLIFERATION OF MOBILE PHONES BEFORE 2005

Firstly, let us examine past developments of the digital music content market in Japan. Japan’s first digital music distribution service started in April 1997. A company called MUSIC.CO.JP began offering songs mainly from independent labels. Coinciding with the launch of numerous music download services in the U.S., a number of online music Web sites opened one after another, orchestrated by individual artists and record labels. In December 1999, SONY Music Entertainment became the first major record company to start an online music store bitmusic. Toshiba EMI, Avex Records, and other major companies followed suit. Yet, since early 2005, the system for online distribution of digital music contents has been underdeveloped, as can be seen in the fact that Mora’s supposed largest online music catalog in Japan contained just 100,000 songs, as opposed to iTunes Music Store’s lineup of 1 million songs upon its launch in August in Japan.

There is no denying that mobile-phone-related music services have been the driving force of the nation’s digital music market. The launch of the i-mode service by NTT DoCoMo in February 1999 marked the start of digital content downloading services via mobile phones. The connection speed of 9600bps in those days made it, initially, difficult to distribute songs in high audio quality. Faced with the adversity, businesses began offering Chaku-melo music ring tones, instead of distributing actual music contents, achieving dramatic growth. The Chaku-melo market has rapidly expanded to 80-90 billion yen in 2002. What makes this development unique was the fact that this service was initiated not by record companies rather by major online Karaoke service providers like GIGA and XING, computer game companies like SEGA, and other companies operating in the peripheral areas of the music industry itself. The market size of 80-90 billion yen as of 2002 is among the highest of all mobile-related digital content services, proving the market-led proliferation of digital content services for mobile phones.

Amidst the flourishing success of the Chaku-melo market, supported by peripheral music
Foreseeing the Future Lifestyle with Digital Music

businesses, record companies that lead the music industry initiated a move to provide the Chaku-uta® service, offering true-tone music as ring tones, instead of Chaku-melo MIDI-based ring tone melodies. The service was initially started solely by Japan’s second largest mobile carrier au in December 2002. While the Chaku-melo service was employed by all mobile carriers rapidly, the Chaku-uta® service was not adopted by the industry leader NTT DoCoMo until February 2004 and by the number three mobile carrier Vodafone until March 2004. However, the service picked up substantial support from younger generations. As the preceding proliferation of the Chaku-melo service had already familiarized mobile phone users with the concept of obtaining music over mobile phone, Chaku-uta® sales reached 100 million songs by July 2004, and surpassed 200 million songs by April 2005 to establish a market of substantial scale. Record companies joined forces to establish Label Mobile, which currently provides around 300,000 songs, approximately three times the catalog size of computer-based online music stores.

After Chaku-uta® came the Chaku-uta Full™ service, which provides whole songs as ring tones to become a de facto digital music store over mobile phones. It reached its fifth million download in April, just 6 months after its launch in October 2004. The cumulative total of downloaded songs reached 10 million in June, causing a dramatic expansion in market size. Although the number of songs available remains smaller than Chaku-uta® at 37,000, the catalog is expected to keep on expanding.

As described thus far, the digital-music-content industry has rapidly mushroomed as one of mobile phone services, but it has been less than 1 year since a full-scale music distribution service (Chaku-uta Full™) was launched. Music has been merely distributed as an additional function to mobile phones, that is, the ring tone. Consumption has been initiated by mobile phone use, instead of music itself. In other words, an explosive proliferation of a new communications device called mobile phones, has triggered the consumption of digital music content as a natural course of evolution. Amidst this situation, a series of dedicated digital music players called iPod has emerged with major success, triggering the launch of the iTunes Music Store offering downloads of digital music content. With the development of a fully fledged environment for computer-based consumption of digital music contents, what has been the course of competition between different types of devices in today’s digital music content market? Let us examine the overview based on the results of a quantitative survey.

TODAY’S DIGITAL MUSIC CONTENT MARKETS: JAPAN AND U.S. MARKETS

In order to grasp the state of today’s digital music content market in Japan, we have simultaneously conducted a survey consisting of identical questions for use in both Japan and the U.S. Results from the two countries were compared against each other in order to identify characteristics of the Japanese market. The survey was titled Survey on Digital Contents, and the survey period ran online was between February and March 2005. The following samples were included: Japan, N=677 aged 15-59 and in the U.S., N=700 aged 18-59.

First, let us take a look at the rate of music-terminal use in the two countries: 6.9% of Japanese respondents used hard-disk music players like iPod, whereas the ratio was almost double at 11.5% in the U.S. The ratio of people using USB flash-memory music players was 7.2% in Japan and 16.1% in the U.S., more than double the Japanese figure. However, the ratio of those using mobile phones as music players was 19.8% in Japan, nearly three times the U.S. result of 6.1%. These figures demonstrated a clear tendency of U.S. users opting for hard-disk or flash-memory...
devices with music transferred via computers, and Japanese users choosing mobile phones to listen to music.

Next, the survey examined how samples typically downloaded digital music contents: 28.9% of U.S. respondents have downloaded music via computer, over 10 percentage points higher than the Japanese ratio of 17.4%. On the other hand, 42.2% have downloaded music (music ring tones) over mobile phones in Japan, around three times the equivalent U.S. figure of 14.2%. The ratio of people who have downloaded true-tone ring tones was 20.4% in Japan, an astonishing lead of around seven fold compared to the U.S. result of 3.0%. The clear tendency of computer orientation in the U.S. and mobile phone orientation in Japan, observed in the choice of music-playing terminals, was also evident in terms of the practice of music downloading.

As explained in the previous section, these findings are a natural outcome reflecting how the digital-music-content market emerged and established itself around mobile phones from the early days in Japan in contrast to market development that evolved around computers and the Internet in the U.S. However, there is some interesting data—the survey asked those who do not own a portable digital music player which type of device they would like to possess. The results indicated almost identical tendencies between Japanese and U.S. respondents, unlike the stark differences they demonstrated in previous questions. Those who intend to purchase a hard-disk music player accounted for 26.7% in Japan and 26.1% in the U.S. The figures for flash-memory players were also very similar at 21.4% in Japan and 21.7% in the U.S. Finally, the ratio of those using a mobile phone as a music player is 5.3% in Japan and 3.0% in the U.S. Even though the Japanese figure is slightly higher than the U.S. figure, they can be viewed as almost at the same level, in comparison to the ratio gaps observed in other questions.

This data demonstrates a strong contrast to previous data, which showed a noticeable computer orientation for the U.S., and mobile phone orientation for Japan. In both countries, purchase intention appears higher for computer-based music players based on either hard disk or USB flash memory, and relatively low for mobile phones doubling as music players.

Until now, Japan’s digital-music-content market has been characterized, in terms of hardware, with proliferated use of mobile phones, as opposed to the U.S. market where more users download music via computer. However, as the results of the aforementioned survey suggest, computer-based music players will be used increasingly for the consumption of digital music content, in addition to mobile phones, in the future Japanese market. Then, what changes will emerge in consumer’s music playing styles when such hardware competition (spread of computer-based music players) evolves?

**FUTURE STYLE OF DIGITAL MUSIC LISTENING IN JAPAN: OVERALL TREND**

We have projected future changes in the style of digital music listening in Japan, dividing the samples of the aforementioned survey into the following three groups and comparing the profiles, music listening styles, and mentalities of the current and future users.

3. Nonusers of digital music contents—Those excluding the above two groups N=319.
We made a particular comparison between current digital music content users who have constituted the computer-oriented digital music market, and digital-music-content potential users who intend to join the market from now, so as to identify how the listening style of this market is likely to change, and what impact such changes will have on the market, which has evolved through downloading by means of mobile phones thus far. First, we compared samples’ demographic profiles and basic indicators in music consumption.

**Gender Comparison**

Current digital-music-content users mainly consist of men, accounting for 66.7%, as opposed to women at 33.3%. Digital-music-content potential users have a more even gender distribution, consisting of men and women at respectively 54.4% and 45.6%. Nonusers of digital music contents have a greater proportion of women at 58.4%, compared to men at 43.6%.

**Comparison by Gender and Generation**

The generation factor was then incorporated to characterize the three groups more clearly. Among current digital-music-content users, men in their 20s claimed the largest proportion at 29.6%, followed by women in their 20s at 23.8%, and men in their 30s at 16.7%. These three groups alone represent over 70% (70.1%), indicating that digital music content is primarily enjoyed among younger people—both men and women. In comparison, among digital-music-content potential users, men in various age groups accounted for around 10% each, that is, men in their 20s at 11.1%, men in their 30s at 10.7%, men in their 40s at 15.3%, and men in their 50s at 11.7%. Women in their 20s and 30s also represented, around the same proportion, at 11.4% and 8.5% respectively. Compared to current digital-music-content users, there is a more even distribution of age and gender groups. As for nonusers of digital music contents, women in the middle to high age groups made up over 40%, including women in their 30s at 14.7%, women in their 40s at 12.3%, and women in their 50s at 16.3%. The data analysis incorporating generation factors highlighted distinctive characteristics among the three user categories.

**Comparison of the Number of CDs Owned**

When asked how many CDs they own, 18.7% of current digital-music-content users said 50 to 99, followed by 23.8% owning 100-199 CDs and 11.9% owning over 200 CDs. These three groups represent over 50% (54.4%). Among digital-music-content potential users, 18.2% own 0-9 CDs, whereas those owning 10-19 CDs, 20-29 CDs, and 30-49 CDs accounted for 14.0%, 13.7%, and 16.3% respectively. Combined, over 60% (62.2%) owned less than 50 CDs. Almost 70% (69.8%) of nonusers of digital music contents also own less than 50 CDs, broken down into those with 0-9 CDs, 10-19 CDs, 20-29 CDs, and 30-49 CDs at respectively 31.7%, 14.7%, 12.2%, and 11.6%. As the figures show, current users have a large proportion of people with a substantial CD collection, whereas nonusers have a large proportion of people with limited CD ownership.

**Comparison of Monthly Music Spending**

Similarly to the former, the ratio of those spending over 3,000 yen (equivalent to the cost of one CD album) per month was 61.8% among current digital-music-content users but less than 40% (39.1%) among digital-music-content potential users. Over 70% (75.9%) of potential users spent at least 1,000 yen (equivalent to the cost of one CD single) per month. Nonusers of digital music contents demonstrated a similar tendency to potential users, with 28.8% spending over 3,000 yen, and 66.1% spending over 1,000 yen. As the
figures indicate, current users have a large proportion of people who spend more on CDs, whereas nonusers have a large proportion of people who spend less on them.

Summarizing the results thus far, current digital-music-content users are mainly young men and women in their 20s, with substantial CD ownership and high music-related spending per month. They can be described as *music fans* with substantial music-related consumption. Potential users of digital music content, who are expected to enter this market, are distributed across both genders and broad generations, from youth to those in middle age. They are characterized as middle-level users in music consumption. Nonusers of digital music content are mainly women in higher age groups, with relative inactiveness in terms of music consumption. The results illustrate clear differences in demographic characteristics and music consumption behavior. There are major differences between consumers who have bolstered the computer-based, digital-music-content market until now, and those who will support the market from now on. These facts alone point to the possibility that the current market is set to undergo substantial changes in its nature. In order to examine details of anticipated changes, we have compared the three groups in their attitude and mentality in listening to music.

### Formats of Music Ownership

Of current digital-music-content users 61.9% acknowledge the desire to store all of their CD collection on the computer, a significantly higher ratio than digital-music-content potential users at 26.7% and nonusers of digital music contents at 17.2%. Current users appear to have a strong desire to manage their music by computer and use computers as the main device for handling music content. In comparison, such desire is not as strong among the other two groups.

### Intention Regarding Songs Available for Downloading

Next, in order to examine the number of songs that are available for downloading, we looked at whether people want a greater selection from download services via computer or those via mobile phone. When asked whether the number of songs downloadable via computer on the Internet should be increased, 45.2% of current digital-music-content users said “yes,” much greater than 30.0% among digital-music-content potential users and 15.0% among nonusers of digital music content. As for whether they want to see the number of songs available via mobile phone increased, just 7.1% of current users agreed, whereas the ratio was more than double at 15.0% among potential users, and 9.7% among nonusers. Although with not as stark a difference as the last paragraph, these groups clearly demonstrated different preferences in catalog enhancement between downloading services via computers or those via mobile phones. In short, current users want to see enhancement of the computer-downloadable catalogs, while potential users want enhancement of mobile-phone-based catalogs just as much as of computer-based catalogs. The results, once again, indicate a strong preference among current users on computer-based services. In comparison, potential users are requesting catalog enhancement to both computer-based and mobile-phone-based services. In other words, potential users and nonusers wish to use both computers and mobile phones to listen to music rather than mere computers.

### Style of Using Songs

We also asked several questions on how people wish to use songs they own. Of current digital-music-content users 35.7% said they want to store all CDs they own on a computer and edit them, for
example, compiling a collection of favorite songs. The ratio was 22.5% among digital-music-content potential users and 11.0% among nonusers of digital music contents. These figures again confirmed the computer-oriented style of current users and highlighted another of their characteristics, that is, actively using downloaded songs for their personal enjoyment. This characteristic became even more evident in the next question.

People were asked whether they like to compile a collection of songs from CDs they own according to specific themes and use the original CD as a gift for friends on a suitable occasion of some sort. Of current users 11.9% said “yes,” whereas the ratio was 15.3% among potential users and 6.0% among nonusers. A greater proportion of potential users expressed preference to this style than current users.

The third question was on whether they wanted to burn their favorite songs on CD-R or DVD more casually to give away to friends and acquaintances. The results showed a similar tendency to the results for the second question. Of current users 7.1% agreed, while the ratio among potential users was greater at 12.7%. Even nonusers had a greater proportion at 9.1%. Looking at the results to these three questions, current users have a self-contained approach in enjoying music with a preference to downloading via computers, whereas potential users are more inclined towards exchanging and distributing music with others.

Finally, we asked whether they wanted to give away or exchange songs, downloaded via mobile phone, to friends over the mobile phone. Again, only 7.1% of current users, who have the preference to computer-based song purchase, agreed to the concept, whereas the ratio was greater among potential users (9.1%), with even nonusers reaching the same level as potential users (8.8%). All the numbers point to the computer-oriented and self-contained nature of current users, and the potential users’ tendency of combined computer and mobile phone use and a strong inclination towards distributing and exchanging songs.

When these analysis results are combined with the demographic characteristics and basic indicators in music consumption, current digital-music-content users can be defined as those with a strong computer preference, wishing to use digital music content for their personal enjoyment in a self-contained approach. In contrast, digital music content, potential users who are entering the market from now, are combining computers and mobile phones for this purpose and are inclined towards enjoying music with others in addition to appreciating it by themselves. This group has a particular tendency of using music as one of the tools for communicating with other people around them.

Let us take a closer look as the results to enable us to explore the direction of how Japan’s digital music market may change, while reflecting upon the trends of both hardware and people’s music listening styles.

The digital-music-content market in Japan originally evolved from the distribution of mobile phone ring tones. Then, music content was merely one of the functions or menus available in using mobile phones. They did not go beyond the ring tone boundary. Amidst this situation, the market embraced the emergence of a new type of music device that contains a hard disk or USB flash memory, designed to be used with a computer. Contemporary music fans were among the first to adhere to such devices, consisting of men and women in their 20s that are most active consumers of music. They stored and managed all music content they already had in a computer, thereby converting them into digital content, and began carrying songs in portable music players and in so doing they were enjoying music for themselves in a rather self-contained fashion.

Today, digital music content that takes the form of mobile phone ring tones exists alongside digital music content that can be carried on hard-disk or USB flash-memory music players. We have investigated the future course of the market in view of the profile and music mentality of potential
users of digital music content, who are making a full-scale entry into this market in the future. Future users will be combining computers and mobile phones, and, unlike contemporary users, enjoying music as both a communication tool with others and for personal entertainment purposes.

Digitizing music contents gives music a new function as a communication promotion factor, in addition to the current functions as ring tones and personal enjoyment.

In order to further clarify this new style of enjoying digital music content, we conducted an oral qualitative survey on two groups, that is, current iPod users who represent those enjoying digital music content via computers, and Chaku-uta® and Chaku-uta Full™ users who represent those enjoying digital music content via mobile phones. The survey clarified their styles in listening to digital music content, so as to obtain an insight into the future direction of music-listening styles.

FUTURE OF THE DIGITAL MUSIC CONTENT MARKET IN JAPAN: CHANGES IN MUSIC CONTENT CONSUMPTION

The Survey on the usage of iPod and Chaku-uta® was conducted in the period between December 2004 and February 2005. In-depth interviews were held with three male and female iPod users in their 20s and 30s, and with three male and female Chaku-uta® Mobile users in their 10s and 20s.²

Comments From iPod Users

What follows are typical comments made by iPod users on their style of enjoying music:

*I find the Random Play function to be very refreshing. I can ‘rediscover’ songs on CDs that I did not pay much attention previously. […] I can store a lot of songs without having to worry about how much space is left. Now, I am storing whatever songs I have, even ones that I would skip if I am playing it on CD. [A 37-year-old man who has used iPod for 1.5 years, SOHO, 6,000 songs are held]*

*I now realize how much I have restricted myself with frameworks of genres and artists when listening to music. The Shuffle function highlights the raw power of individual songs. […] Occasions like that have broadened the range of music genres I listen to, making me feel like trying out CDs I would never have dreamed of listening to before. [26-year-old woman who has used iPod for 1 year, office worker, 1,500 songs are held]*

*I never used to carry music around, but now I do not go anywhere without my iPod. This has widened the variety of occasions I listen to music, for example, while on a train or on the way home after a drink. […] There are times when a song I frequently listened to on CD sounds very different on a portable player, because of various situations you are in at the time. That gives me fun. [39-year-old man who has used iPod for 6 months, office worker, 700 songs are held]*

As suggested in the results of the quantitative survey, they typically—to some extent—have a self-contained approach in music entertainment. Their remarks illustrate new ways of enjoying music (consumption styles) they have attained through hard-disk music players like iPod. For them, a hard-disk music player is a device that allows them to randomly enjoy music out of a greater selection of songs than previously possible in conventional devices (cassette player, MD player), loaded from their CD collection. A hard-disk music player is a true portable music player strictly for personal use. In order for the device to be self-contained, it must be able to carry a massive number of songs, which in turn, facilitates random playing. This random playing then releases listeners from the boundaries of
Forseeing the Future Lifestyle with Digital Music

existing music context (by genre, by artist, etc.) and invites the creation of new contexts, thereby enhancing the self-contained nature even further. They are enjoying music in this cycle.

For them, consumption of music content is not about listening to each individual song, but about enjoying a stream of music. It must always be a fresh string or stream of music different from what they have already experienced previously. Their consumption of digital music content is characterized as self-contained context consumption. This is an emergence of a new style of music consumption, only possible for hard-disk music players like iPod. The style is facilitated with the concept of *play list* in iPod and other devices. The ability to compile a new string or stream of music, has diluted the concept of *album*, presented conventionally from package providers and artists as producers, and encouraged individual users to compile their own music streams. Consequently, music is increasingly evaluated on the merit of each song. One example is the way iTunes Music Store presented its proliferation scale in the unit of individual songs downloaded, rather than albums. This kind of presentation appears to depict a transition of mentality, with listeners focusing more on individual songs, rather than embracing the supplier-defined unit of *album*.

The following comments are derived from Chaku-uta® Mobile users on their style of enjoying music:

*During a break time at work, I just leave my mobile phone to play songs to provide some background music. [...] They are all songs that everyone knows, and will not trigger any music talk. However, it is better than having no music, and stimulates conversation. [...] I don’t mind if each song may be just 30 seconds long. It is actually better to have short tunes to enjoy them with my colleagues and build up a lively atmosphere.* [25-year-old man who has used Chaku-uta® Mobile for 6 months, office worker, 10 songs are held]

*I like upbeat Techno music. I use these types of songs as Chaku-uta, and play them during break time when my friends are around, so that I can show them how to dance to the tunes. [...] The other day, I had my mobile hanging around my neck, and playing Chaku-uta, as I danced across a Shinjuku intersection with my friend. [21-year-old woman who has used Chaku-uta® Mobile for 6 months, university student, three songs are held]

*I might listen to and check out music with my friends, but I am more likely to use it as the Sleep Timer when I go to bed. [...] I can operate it by hand, and put it on charge at the same time. It is very convenient. [...] I don’t care about (each song being the ring tone length of 30 seconds and) not having the complete song. I fall asleep as I listen to the same songs repeatedly.* [19-year-old woman who has used Chaku-uta® Mobile for 1 year, vocational school student, five songs are held]

The analysis of the quantitative survey results also indicated that persons entering the digital-music-content market from now use both computer-based music players and mobile phones, and use music to enhance their relationship or communication with their friends and acquaintances instead of merely enjoying music by themselves. Comments from Chaku-uta® Mobile users substantiate the tendency, that is, using music as a tool for sharing various occasions with friends.

As part of the quantitative survey described earlier, people were asked how they use ring tones and Chaku-uta® on mobile phones. The top three answers were as “ring tones (87.3%),” “alarm clock (60.4%),” and “other alarm sounds (46.3%).” The fourth highest ranked answer, however, was to “enjoy them alone as music” (44.0%), and 41.7% said they “enjoy them together with friends or use them to entertain others”—as such indicating that people are beginning to enjoy mobile-downloaded tunes as stand alone songs with friends.

What can be noted in these comments is that songs are enjoyed in the ring tone length of 30
seconds, rather than in their entirety, which is quite different to that of hard-disk music player users, who consume a massive amount of randomly replayed music in various contexts. Their consumption style is summarized as “sharing the occasion of playing popular songs, rather than personal favorites with others to magnify enjoyment.” What counts is how good each NETA (=song as conversation topic) is, rather than how many songs you have in store. Their consumption of digital music content is characterized as “NETA consumption to be shared among friends.” For them, Chaku-uta® Mobile is perceived as a “music player that plays 30-seconds of everyone’s favorite songs for a shared experience.” The emergence of a new service called Chaku-uta® has brought about this new style in music consumption, while now, the style seems to transform formats of music content.

CONCLUSION

As we have examined, Japan’s digital-music-content market—which started off with the distribution of ring tones as one mobile phone service—has embraced the arrival of fully fledged digital music players and online stores, both designed to be used via computers, such as iPod and iTunes Music Store. From the viewpoint of hardware competition, the market has now entered a stage of combined development of mobile-phone-based devices and computer-based devices. It has brought about two contrasting consumption styles with distinctive characteristics (computer-based and mobile-phone-based consumption of digital music content), and diversified people’s styles in enjoying music at the same time.

People who use a computer-based means to enjoy digital music content, have a self-contained style of consuming music in a specific context, loading a hard-disk music player with a greater amount of music from their personal CD collection than previously possible and enjoying songs in random order. In contrast, people who use mobile-phone-based devices, employ a mobile phone as a communal music player for playing 30-second tunes of high popularity and consume music as topics (information) for sharing various occasions with friends or enhancing the atmosphere.

At present, these styles are separate tendencies and can be observed among users of hard-disk music players and users of mobile phones as music players as two extremes. However, a steady proliferation of hard-disk or USB flash-memory music players may cause these styles to merge on the side of individual users. Competition between two types of devices has created two distinctive styles of listening to music. Now, each user may start using both of these devices at the same time, hence adopting both styles alongside each other. Such a user may eventually begin to seek both of the styles in one of the two types of devices, which may amount to hardware integration, brought about by the symbiosis of the two different music-listening styles. Closely paying attention to consumer behavior and practices in the future will then give way to rich empirical data to be used to develop and elaborate the stream of thought outlined in this study further.

FURTHER READING


Foreseeing the Future Lifestyle with Digital Music


**ENDNOTES**

1 The survey was conducted by Macromill Inc. in Japan and Zoomerang, Inc. in the U.S. And, it was organized by Hakuhodo Institute of Life and Living and Hakuhodo DY Media Partners’ Media Environment Laboratory.

2 The survey was conducted by Oval Communication and was organized by Hakuhodo Institute of Life and Living and Hakuhodo DY Media partners’ Media Environment Laboratory.

This work was previously published in *Information Communication Technologies and Emerging Business Strategies*, edited by S. van der Graaf, pp. 59-75, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global)
Chapter 3.15
Widely Usable User Interfaces on Mobile Devices with RFID

Francesco Bellotti
University of Genoa, Italy

Riccardo Berta
University of Genoa, Italy

Alessandro De Gloria
University of Genoa, Italy

Massimiliano Margarone
University of Genoa, Italy

ABSTRACT
Diffusion of radio frequency identification (RFID) promises to boost the added value of assistive technologies for mobile users. Visually impaired people may benefit from RFID-based applications that support users in maintaining “spatial orientation” (Mann, 2004) through provision of information on where they are, and a description of what lies in their surroundings. To investigate this issue, we have integrated our development tool for mobile device, (namely: MADE, Bellotti, Berta, De Gloria, & Margarone, 2003), with a complete support for RFID tag detection, and implemented an RFID-enabled location-aware tour-guide. We have evaluated the guide in an ecological context (fully operational application, real users, real context of use (Abowd & Mynatt, 2000)) during the EuroFlora 2006 international exhibition (EuroFlora). In this chapter, we describe the MADE enhancement to support RFID-based applications, present the main concepts of the interaction modalities we have designed in order to support visually impaired users, and discuss results from our field experience.

INTRODUCTION
Starting from the European Union cofounded E-Tour project, we designed the tourist digital assistant (TDA) concept and developed multimedia
widely usable user interfaces on mobile devices with RFID

tour guides on mobile devices (PocketPC and Smartphone devices) for a number of European tourist sites, such as the Costa Aquarium of Genoa, “Strada Nuova” architectonical area and the city of Genoa, the Castellon region in Spain, and the city of Uddevalla in Sweden (Bellotti, Berta, De Gloria, & Margarone, 2002).

The tour guide provides multimedia contents, added-value information, and location-based services to the tourists. Added-value services are implemented by integrating the mobile devices with additional hardware and software tools such as GPS, electronic compasses, wireless connectivity, digital cameras, written text input, databases, and so forth.

See Figure 1 for snapshots of tourist guide applications.

Relying on the argument that “play is a powerful mediator for learning throughout a person’s life,” we developed the “educational territorial-gaming” concept in VeGame (Bellotti, Berta, De Gloria, Ferretti, & Margarone, 2003), a computer-supported educational wireless team-game played along Venice’s narrow streets to discover the art and the history of the city (see Figure 2), and in ScienceGame (Bellotti, Berta, De Gloria, Ferretti, & Margarone, 2004), a sort of treasure-hunt game inviting players to discover the mysteries and the marvels of the science (see Figure 3) during the “Festival della Scienza” exhibition held in Genoa every year.

These applications were developed from scratch. From these first experiences, we identified common needs and came up with a system to support design of multimedia applications for mobile devices, called Mobile Applications Development Environment (MADE) (Bellotti et al., 2002).

MADE includes M3P (MicroMultiMedia Player), a network-enabled multimedia player easily programmable through the micromultimedia services language (MSL). MSL provides high-level components encapsulating advanced services (e.g., positioning, database query, path search, etc.) that can be easily integrated in multimedia applications. This allows building modular software programs that provide information-rich services to the general public through a coherent and homogeneous HCI that can be learned with low mental workload. On the other hand, MADE hides the low-level aspects of multimedia and service management, allowing designers to focus on the modalities of presentation of information and on user interaction, reducing learning, development, and code maintenance time.

In this chapter, we describe the latest MADE enhancement: we have integrated it with a complete support for RFID detection to allow develop-

Figure 1. Snapshots from the Aquarium and Strada Nuova tour guides on PocketPC device
ment of multimedia mobile applications directly connected with the physical world (Want, Fishkin, Gujar, & Harrison, 1999). All low-level aspects of the hardware tag-detection system that are necessary to identify and locate physical objects with attached small RF tags (Want, 2004) are hidden to MSL programmer by the MADE system.

This chapter will also show the use of MADE with the RFID support in a real context such as EuroFlora 2006 international exhibition. This guide differs from others because it has been ad-hoc developed in order to meet strict usability needs. In particular, the novel interface design assists visually impaired people in maintaining “spatial orientation” (Mann, 2004) through provision of information on where they are, hazards that might be in the way, and a description of what lies in their surroundings.
MADE SUPPORT OF RFID TECHNOLOGY

Location-Aware Computing

Recent research has developed several systems, to determinate physical location, that differ by accuracy, cost, and coverage (Boriello, Chalmers, LaMarca, & Nixon, 2005). The global positioning system (GPS), which uses signal from satellite to estimate position (Djuknic & Richton, 2001), is the most used system, but only for applications in outdoor areas. In indoor areas and urban areas with poor sky visibility, the system does not work properly. Moreover, it has a long start-up time.

To overcome these limitations, the first indoor positioning system was the active badge system (Want, Hopper, Falcão, & Gibbons, 1992), which is based on sensors that receive infrared ID broadcast from tags worn by people. This system gives a poor (room-grained) localization precision. After the active badge system, typical indoor location systems are based on radio frequency and on the estimation of position computed from the measured signal strength. Various technologies can be used: Wi-Fi (Howard, Siddiqi, & Sukhatme, 2003), Bluetooth (Bruno & Delmastro, 2003) and nowadays RFID (Liu, Corner, & Shenoy, 2006).

The first two solutions can give an accuracy of around some meters, but require expensive fixed base stations. RFID tags, instead, are very inexpensive and have the same performance. The literature reports also of many location estimation algorithms based on cellular radio networks (Xu & Jacobsen, 2005). However, there is not a generally agreed solution today, and each algorithm has pros and cons, depending on environmental issues. Finally, some vision-based algorithms (López de Ipiña, Mendonça, & Hopper, 2002) are promising because they do not require infrastructure (like tags, satellite, or base station). However, it is difficult to set up a system to locate a user with a 1-meter precision. In the selection of the best methodology for our system, we have taken into account three major issues: the possibility to have a system for outdoor/indoor spaces (like the EuroFlora 2006 exhibition area), a technology with a low cost for the deployment of the infrastructure, and a likely pervasive availability of the system in the near future. All these requirements are satisfied by the RFID technology.

RFID Application Fields

Major RFID application domains include monitoring physical parameters, such as temperature or acceleration, during fragile or sensitive products delivery, monitoring product integrity from factory to retail locations (Siegemund & Floerkemeier, 2003), utilities for home and office automation (Langheinrich, Mattern, Romer & Vogt, 2000). Nowadays we have passive or active inexpensive RFID (approaching 35 cents today, with a goal of 5 cents (Quaadgras, 2005)) that makes these kinds of sensors practical for tourist applications. For example, a museum exposition can place tags attached to each point of interest so that tourists can receive information about exposition in the right moment at the right place; when near to the object. The research community has actively explored this possibility at the Exploratorium, the interactive science museum in San Francisco. The HP Laboratories researchers have implemented a system that uses three types of identification technology: infrared beacon, barcodes, and RFIDs (Fleck, Frid, Kindberg, O’Brian-Strain, Rajani, & Spasojevic, 2002). In Goker et al. (Goker, Watt, Myrhaug Whitehead, Yakici, Bierig, et al., 2004), a special tag that can work with mobile devices to provide ambient information to users on the move is described. In the Cooltown project (Kindberg & Barton, 2001), RFIDs are used to attach pointers from everyday objects to entities in the computational world. A full exploitation of RFID potentials requires study and implementation of human–computer interaction (HCI) modalities able to support usability of the enhanced mobile tool by the general public. This implies the neces-
usability to resort to programming methodologies and tools specifically dedicated to support the RFID technology. Thus, we have extended the MADE toolkit to support a link between applications and physical world through RFID sensors.

**MADE Architecture**

A typical MADE application consists of a set of pages containing multimedia and service objects. The micromultimedia services language (MSL) script specifies pages’ layout and objects’ appearance, synchronization, and user-interaction modalities. MSL scripts are interpreted at runtime by the M3P player that manages presentation of contents and user interaction according to the instructions specified in the input MSL script.

M3P player relies on two-layer architecture (see Figure 4) involving a high-level, platform-independent director and a low-level driver. The director is responsible for creating, initializing, and managing the objects that implement the language functionalities. In order to support incremental development of the player, M3P is composed by a set of modules. In particular, the director has been designed to be independent of the components it manages. According to the object-oriented methodology, this has been achieved by encapsulating the functions of the components in the code of their class, and by structuring the hierarchy so that the director can simply keep a reference to the presentation’s pages and convey them events.

According to the instructions specified by the MSL developer in the script, events (either from the system or from user interaction) are conveyed to the director that simply redirects them to the components of the page currently on show, which is the higher-priority choice or, with lower priority, to the other pages of the presentation.

Events are implemented as string messages that are to be interpreted by the target objects. This design choice allows the director’s code to be independent of the components and the components to be independent of each other. The basic assumption of this schema is that each component exports a well-defined interface (i.e., a set of messages to which it is able to react) and implements this interface (i.e., implements the reaction to such messages).

![Figure 4. MADE architecture](image-url)
Thus, components can be seamlessly added and interchanged (in this last case, as long as they comply with the same interface). Adding a new component (i.e., a new functionality) does not involve any change either in the director’s code, or in the other components’ code.

Such a design choice supports easy incremental development, allowing seamless integration of services within a single application framework. This implies that a homogeneous HCI can be applied to an application that hosts several different services that have been developed independently of each other (e.g., intelligent tour planning, interactive maps, positioning, and database access).

MSL relies on a component-based data structure. That is, an MSL file specifies creation of components, attributes of components, and their reaction as a consequence of user interaction. Components are organized in three main libraries: multimedia (e.g., audio, image, video, button), synchronization (utilities like timers that can be used to implement synchronization and scheduling of contents), and services (objects that encapsulate services such as positioning, shortest path search, tour planning, database query, etc).

Every different type of component has its own kind of attributes (fields). The fields record data for specifying the appearance (such as position) and the behaviour (i.e., reactions to events) of the component. In general, components are contained in a special container component, called CARD, that can be thought of as an empty page on which the developer can add components.

The core of M3P involves a platform-independent director and a platform-dependent driver. The director manages the multimedia objects that implement the presentation. Objects are compounded in hierarchical structures. For instance, a CARD (i.e., a multimedia page) may include several images, buttons, and mpeg players. The driver implements the functions to access the hardware, while the director deals with the logic of the multimedia presentation.

### Integration of RFID Subsystem

A major feature of MADE consists in the possibility of incrementally adding new hardware and software modules, that are integrated into the HCI framework with no need for modifying the M3P core, since every component’s class is responsible for interpreting its receivable messages, independent of the others. MADE can integrate, into a common framework, various hardware modules independently developed to augment the mobile device potentiality. M3P driver’s classes, which have to be developed to integrate every new hardware subsystem, manage low-level aspects of the hardware modules, while the MSL interface to the application developer abstracts the services at high level. This implies that a homogeneous HCI can be applied to an application that hosts several different services that have been developed independently of each other (e.g., automatic positioning, intelligent tour planning, and database access can be integrated in an interactive map), and the MSL developer can simply exploit the service modules focusing on the integration of the HCI. Examples of hardware modules already integrated in MADE are a positioning and orientation module that an MSL developer can exploit to get geographical position from a GPS receiver and direction from a digital compass, and the remote communication module able to exploit the hardware available for connection with the external world (e.g., wired/wireless LAN, Bluetooth, GSM/GPRS cellular networks).

In order to enable applications to react to objects in the physical world, our new M3P module, called RFID sensing module (RfidSM), detects presence of RFID tags in the surrounding space and notifies the M3P run-time objects with events implemented as string messages (see Figure 5). The script interface of the RfidSM is a new MSL component, called RFID, that exposes the fields shown in Table 1.
When the RfidSM component is started, and until it receives a stop event, it scans the surrounding environment to check the presence of tags every “period” of time. The list of detected tags is then sent with the MADE message-exchange modalities to the components specified in the

<table>
<thead>
<tr>
<th>Component Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>List of identifiers of the components to which information about identified tags are sent</td>
</tr>
<tr>
<td>Period</td>
<td>A time period in milliseconds between two consecutive environmental scans to detect tags</td>
</tr>
<tr>
<td>Repetition</td>
<td>A number of tag detection operations executed consecutively on each scanning action</td>
</tr>
<tr>
<td>Id</td>
<td>A list of RFID tags that are of interest for the component</td>
</tr>
<tr>
<td>Delay</td>
<td>A list of time frames, one for each interesting tag, in which tags are not identified again</td>
</tr>
<tr>
<td>dBm</td>
<td>A list of signal strength values, one for each interesting tag, that specify thresholds for tag identifications</td>
</tr>
<tr>
<td>onFound</td>
<td>A list of events, one for each interesting tag, that RFID component launch when a tag is identified</td>
</tr>
<tr>
<td>Start</td>
<td>If a component launch this event on a RFID component starts the scanning of tags</td>
</tr>
<tr>
<td>Stop</td>
<td>If a component launch this event on a RFID component stops the scanning of tags</td>
</tr>
</tbody>
</table>

Table 1. RfidSM fields description
“target” field. In addition, the component has an “id” field to allow programmer expressing interest in a set of tags, and defining (through the field “onFound”) the corresponding events list that should be executed. Each interesting tag is also featured with a signal strength threshold (through the field “dBm”) that specifies a limit under which the tag is considered in range.

There is the problem of collisions, since the scan results are typically imperfect due to not all tags are detected in every scan. To solve this problem, a tag typically awaits a random number of time slots before it answers the RF pulse sent by the reader. However, the problem still remains and grows as the number of tags in the surrounding environment grows. The MADE RFID sensing module tackles this issue, allowing the programmer to specify, through the field “repetition,” a number of times that the reader should repeat the scanning before returning to the founded tags. The list of founded tags is the collection of all tags observed in each scan. A small value of repetition results in a fast scan with high risk of collision, whereas large repetition value results in a slow scan with few collisions. This trade-off should be resolved by the programmer basing his decision on application constrains: long delays can result in human-computer interaction problems if the application allows a user expectation for immediate reaction to tags. This is the case of applications in which user a voluntarily accosts the mobile device to tagged objects to obtain information. Similar problems arise if the application has a short time frame to detect tags, for example, in applications where the user moves at relatively high speed in the environment, like in territorial games. Instead, others type of applications can gain advantage from precise but slow detections. It is the case of a tourist mobile guide for a naturalistic park in which the user moves along a path with some points of interest largely spaced, like tree species or rare flowers.

The other problem affecting the RFID technology is the “tag detection flickering” (Römer, Schoch, & Mattern, & Dubendorfer, 2003): due to the collision problem, some tags can appear and disappear from sequential scanning, generating a fast list of tag identifications. The MADE RFID sensing module allows the programmer to decide how to convert scan results into applications events handling this problem. Programmer can specify, through the “delay” field, a time period (for each interesting tag) starting from the detection. During this time, subsequent detection events of the same tag are discarded; also, the exact definition of this delay is application dependent. Applications with events that occur only one time, like tourist guide for museums with linear path, can have delay values set to infinite. Instead, in applications with events generated multiple times closer each other, like territorial games, the delay should be short or zero.

Currently, we have implemented the low-level driver support for the iCARD Identec reader in a PCMCIA card format (IDENTEC). This card can be integrated in handheld, portable, or laptop computers to communicate with the iQ and iD active RFID tags at a distance of up to 100 meters. The RF signal is in the UHF radio band (915 MHz or 868 MHz), providing long-range communication and high-speed transmission rates for reliable data exchange.

THE EUROFLORA GUIDE

In order to assess the possibility of using RFID technology to develop widely usable interfaces, we present a real-world application (see Figure 6) developed through the MADE toolkit and deployed in an ecological environment (fully operational, reliable, and robust application, used by real users and in a real context of use) at EuroFlora 2006 (the international flower exhibition that is held in Genoa every 5 years). With over 500,000 visitors in 10 days, EuroFlora is one of the most important exhibitions of Europe.
The developed application concerns the research area of assistive technologies for visually impaired people. Such assistive applications have the potential to improve the quality of life of a large portion of population (by 2020, there will be approximately 54 million of blind persons over age 60 worldwide (WHO, 1997)).

In this field, maintaining spatial orientation is a major challenge for people with visual impairment. There is the need of systems in providing blind people with information on where they are, hazards that might be in the way, and a description of what lies in their surroundings (Mann, 2004). The notion of “spatial orientation” refers to the ability to establish awareness of space position relative to landmarks in the surrounding environment (Guth & Rieser, 1997). The goal of our application is to support functional independence to visually impaired people, providing support to indoor awareness of elements in the surroundings (Ross, 2004).

The EuroFlora guide is organized in two parts. One part provides general information about the exhibition, the guide, and their services. The other part provides the description of the selected interest points. While first part is directly accessible by the user at any moment, the second one is event driven. More precisely, every interest point description is associated to an RFID tag, and when a user enters that area (i.e., her/his handheld device recognizes the RFID tag), the software asks the user whether to launch the corresponding description.

We placed 99 RFID sensors on an area of 30,000 mq of exhibition, covering 99 points of...
interest, services, and major areas (see Figure 7). RFID sensors were IP65 compliant in order to resist to water and dust, and self-powered. Power level of sensors could be set in two levels, low and high.

**Design Methodology**

The necessity for combining the flexibility and multimedia potential of a mobile device with the extreme simplicity of interaction, required for use by a wide audience (also visually impaired people), involves facing three main HCI issues:

- **Usability by general users:** The tourist has little time and willingness to learn how to use the new technological tool, since she or he is there to visit the exhibition and not to learn a tool. Most of the tourists use such a tool for the first time and just for a short time (typically, from 30 to 90 minutes). Thus, the complexity of the platform should be hidden from visitors, making the guide immediately usable, with no effort by users. This implies that the interface is to be as simple and intuitive as possible.

- **Usability by visually impaired people:** Visiting an exhibition is a critical task for the blind, mainly for the combination of several reasons: the site is often crowded and unfamiliar to the visitor, it may be noisy, it is difficult to orientate in a highly dynamic place. In this context, the guide should be not intrusive, with few and very recognizable input interface elements (also with tactile feedback), and should give information in a proactive modality when needed by the user.

- **Presentation of information:** Added-value information (e.g., how the various specimens live in their natural environment) should be synergistic with the direct experience of the visitor at the exhibition. Provision of information has to be structured in order to enhance the direct perception of the visitor, leading to a better and more pleasant comprehension of her/his surrounding environment. For example, the guide should make use of environmental sound (e.g., waterfall) and scent (e.g., flower smell) to connect content information and the objects in the space.

We have tackled such issues resorting to the methodologies of the user-centric design (Carroll, 1997), in an iterative development of the guide involving participatory design (Beck, 1993), definition of usability specifications, and contextual design, as shown in the following:

- Participatory design consisted of the participation of botanists at the design decisions, authors skilled in writing for blind people and visually impaired end-users, together with technical developers. The most significant contribution of the first three categories consisted in the definition of the targets and in the concrete perspective they brought into the project.

- Usability specifications provide explicit and measurable targets to verify the suitability of the work done. Examples of such goals are “90% of the users should be able to operate the guide without asking questions to the personnel,” “90% of the users should be able to use the interface with no errors,” “90% of the users should be able to understand the meaning of all the touchable controls within 120 seconds.” All these objectives were verified in early lab and field tests in order to take the appropriate corrective actions.

- Contextual design involved early field tests with experts and users at the exhibition in the preopening days when the infrastructure of EuroFlora was being built. Field tests have been helpful to highlight problems and shortcomings that had been overlooked or ignored in lab.
Structure of the Interface

The interface of the EuroFlora guide has been designed to support immediate use by the general public, also by visually impaired people. To this end, we used general design principles (we already described them in the introduction) such as overall simplicity, low intrusiveness, and support for natural interaction and knowledge acquisition. Moreover, we added further features in order to meet the specific needs of visually impaired people:

- Tactile feedback in the control interface
- Tutorial stage
- Event-driven interface
- Facilities to support orientation

The basic element of the interface is the multimedia card. A multimedia card corresponds to each subject of a presentation (e.g., a flower species). Each multimedia card provides, in an audio format, texts specifically written for visually impaired people (i.e., highlighting olfactive and tactile sensorial information, providing detailed ambient descriptions).

The tactile feedback is necessary to allow impaired people to easily understand the position of the controls and give her/him feedback. Our previous multimedia guides had the interface embedded in the graphic contents, exploiting the touch screen of a pocket-pc device. During the early field tests, visually impaired people pointed out some important shortcomings in these solutions. They felt that the screen was too large and their fingers were lost in a space without roughness. Since most of such users are well acquainted with the common cell phones’ relief keyboard, we studied a new solution exclusively based on the hardware buttons (Figure 8).

The hardware buttons of the interface are highlighted. The choice of this area as navigation control allows visually impaired people to have a tactile feedback. The meaning of the button is “up” to accept a content description (which is automatically triggered when the user enters a new cell), “down” to reject it, “right” to exit from a section of the guide, and back to the main menu, “left” to have contextual information about user’s current position.

The tutorial stage is an initial guide section in which users could freely experiment with the interface of the tool in order to allow people to use the guide in an independent way. In this stage, users are invited to freely press buttons. A speech description briefly explains the meaning of each pressed button. This tutorial stage prevents the necessity for providing papers or long explanations when users rent the guide.

The event-driven interface allows a user to get information about points of interest (POIs) and orientation when they are in the proximity of a POI. For example, in Figure 9, a user near the azalea stand is told about the presence of this flower by a pop-up window (the guide is to be usable by everybody) and a corresponding jingle sound. If she/he wants to listen to this content, she/he can press the “up” hardware button. By
default, the system skips the presentation. This operational mode has low intrusiveness (users are asked whether to listen to a content), but it also provides a certain degree of proactivity. Information is not only botanical, as in the example of azalea, but also concerns the positioning of the user. Many tags are placed in the proximity of facilities, such as lavatories, cafés, exits, and intersections. This localization system lowers the mental workload necessary for the tourist to synchronize the physical space of the exhibition with the virtual space of the program.

This message (accompanied by a jingle) is shown to the user when she/he has just entered a POI area. The combination of audio and graphics is due to the fact that the guide may be used also by not visually impaired people. In the example in this figure, the user is near to the azalea flower, and if she/he is interested in the description she/he can press the “up” hardware button to access the related content.

One of the main tasks of the guide is to assist the visitor in her/his exploration of the exhibition space. A facility to support orientation (not useful for visually impaired people) is a section with the map that helps the tourist to orient herself/himself in the exhibition. The map (see Figure 10) shows the structure of the EuroFlora, including lavatories, café, exits, and so forth, and the location of the points of interests. In order to enhance user’s orientation, the map is centered on the position of the user, as determined by the currently perceived RFID tags.

FIELD EVALUATION

Experimental Framework

Real evaluation of advanced mobile device applications and of the impact on their intended population is difficult and costly. Evaluation requires analysis of real users, in a real context of use. In order to adequately evaluate interaction with computation resources, test-people should use a fully operational, reliable, and robust tool, not just a demonstration prototype (Abowd & Mynatt, 2000). Hence, it is important to perform
early tests in the authentic context of use in order to verify end-user acceptance and overall usefulness of the system, and to receive feedback to inform future design.

In this chapter, we describe the early stage analysis of acceptance and usefulness of the developed multimedia guide. The tests were performed at EuroFlora 2006, the international flower exhibition that is held in Genoa every 5 years. With over 500,000 visitors in 10 days, EuroFlora is one of the most important exhibitions of Europe.

The exhibition area (around 90,000 squared metres) was equipped with an infrastructure of 99 RFID tags. The experimentation involved 120 visually impaired tourists who used the tour guide, and were observed and interviewed by expert of disabilities and HCI designers. Subjects were visually impaired (25%) or blind persons (75%) representing a wide range of age (28% age <30; 32% age between 30 and 50; and age >50 40%). Moreover, the tests involve 64 females and 56 males.

The tour guide consisted of a PocketPC equipped with a special leather package with a lace dropping from the neck for a more comfortable use. Headphones were also used in order to isolate the user from the highly noisy surrounding environment (see Figure 1).

Preexhibition Tests

In an early test session-performed 2 days before the official opening of the exhibition, when some stands were already ready enabling a realistic test-we prepared a prototype software version that was used by five selected visually impaired users visiting 30% of the total exhibition area. We followed and interviewed the users in this phase, in order to understand shortcomings, defects and weaknesses, and strong points of the product. In this phase, we understood and solved some problems on user interface and contents, such as the most suited assignment of buttons to presentation control functionalities and the length of the descriptions. Some test-users found the long silence time between a presentation activation and the next one (i.e., the period of time in which the user is walking through areas not covered by RFID tags) frustrating. We partially tackled this issue by periodically providing a message saying that the user is currently in an area not close to a POI.

Ecological Tests

One hundred and twenty blind people used the guide during the exhibition. Sixty of them (aged from 12 to 78 years old) participated in

<table>
<thead>
<tr>
<th>Issue</th>
<th>Average</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usability</td>
<td>4.00</td>
<td>0.64</td>
</tr>
<tr>
<td>Usefulness</td>
<td>4.25</td>
<td>0.75</td>
</tr>
<tr>
<td>Support for spatial orientation</td>
<td>4.20</td>
<td>0.66</td>
</tr>
<tr>
<td>Session length time</td>
<td>201 minutes</td>
<td>30 minutes</td>
</tr>
</tbody>
</table>
an ecological test conducted by the authors. We interviewed the users at the return of the guide. We evaluated three main performance factors: usability (including effectiveness, efficiency and pleasantness of use), usefulness, and capability to support spatial orientation (in particular the approach to the POIs). We asked users to give a general comment on the guide and a 1-5 grade for each factor (which was carefully explained and defined by the interviewers). An overall survey of results is reported in Table 2; it clearly shows the high acceptance by the users.

Analyzing the variables' correlations based on the chi-square test, we observed that usability is correlated with the perceived support for spatial orientation ($\chi=25.3$, df (degree of freedom) $= 16$, 90% confidence), and that perceived utility of the tools is strictly correlated with perceived support for spatial orientation ($\chi=30.2$, df =16, 99.9% confidence). This suggests the importance of our design choice to use mobile technology to support orientation of visually impaired people. Moreover, test results also show that the tool is perceived as useful and usable.

Considering the free comments, the guide was judged as an excellent tool for users to orientate themselves inside the exhibition. Several people expressed a similar concept, which we can synthesize with the words of one visitor: “after always having been guided, for the first time I myself have been able to guide my wife and to explain the exhibition!” Such positive comments were also confirmed by the blind assistance experts, who highlighted the significant degree of independence the blind could reach through the guide.

Shortcomings in the interface were reported by some elderly users, while some people asked for more extended descriptions, though each point of interest included at least one. The high performance and reliability of hardware, software, and batteries assured long sessions of use with no troubles for the user.

**FUTURE TRENDS AND VISION**

The research community is envisaging a new model of a “tagged world” as an intelligent environment that allows providing visually impaired people with information about architectonical barriers, safe paths, points of interest, potential danger areas, and other useful information. A sample scenario description may give an idea of this likely future.

Maria is visually impaired. She is in a foreign city on a business trip. Maria owns a mobile device with a mobility-assistance system (MAS: it is similar to the EuroFlora Guide, but with a much larger action range). The MAS accompanies her in her path to her destination office, and signals pedestrian crossings, traffic lights, safe paths in work-in-progress areas, and so forth. All objects in the world send their signals, but Maria’s wearable device has an intelligent reasoning algorithm (based on user preferences and interpretation of the user’s current activity) and a suitable human-computer interaction (HCI) in order to provide her only with the needed information. This information is extracted from a mass of data that are continuously received from the close-by RFID tags. Thus, the wearable device notifies Maria about a pedestrian crossing only if it knows that this is useful for her current activity (i.e., going to office). Not useful information will not be provided, in order not to distract Maria. Along her path to her destination, Maria passes by a newsagent. The World Guide scans all the magazines and identifies today’s issue of Maria’s favourite magazine. It queries Maria’s database, which replies that Maria has not purchased this issue yet; so, it notifies her about the opportunity to buy the magazine.

**CONCLUSION**

The ubiquitous presence of smart tags will offer, in the near future, a critical mass of information,
Widely Usable User Interfaces on Mobile Devices with RFID

embedded in the world, that will be exploitable to rethink the relationships between people involved in their daily-life activities and the surrounding world.

With MADE we have designed a system that continuously scans the tagged world, interprets the large amount of information coming from the surrounding objects, and provides it to the users through multimedia human-computer interaction. Moreover, the application in the future will filter the raw data coming from the environment (with artificial intelligence behaviour) taking into account the user needs, preferences, and profile.

The field test at EuroFlora 2006 has demonstrated the feasibility of our vision, by deploying the system in a real-world setting (an exhibition area with indoor and outdoor instrumented environments), and performing extensive field tests with real users. In a longer-term view, with such an application, we intend to investigate the future scenarios that will be enabled by a massive presence of RFID tags in our environments. This “early prototyping” has allowed us to understand, as early as possible, costs, limits, strengths, and benefits of the new technology. We have also obtained a significant positive feedback on user acceptance. Usability results show that the guide is perceived as highly usable and useful, in particular because of its ability to support spatial orientation.

The next step towards a “tagged world” will require integration of data and services, and capability of interpreting a variety of sources according to the specific and dynamic user needs. Achieving these goals will involve a huge research effort that will be successful only if it will lead to the deployment of compelling applications that will be perceived as useful by the users. In a user-centered design view, this implies a rapid prototyping of applications and extensive user testing in the real context of use, which was our inspiring principle in the EuroFlora project.

REFERENCES


Carroll, J. M. (1997). Human-computer interaction: Psychology as a science of design. *Inter-
Widely Usable User Interfaces on Mobile Devices with RFID

national Journal of Human-Computer Studies, 46(4) 501-522.


KEY TERMS

Chi-Square Test: The Chi-square is a test of statistical significance for bivariate tabular analysis (crossbreaks). This test provides the degree of confidence we can have in accepting or rejecting a hypothesis.

Ecological Context: The ecological context is a set of conditions for a user test experiment that gives it a degree of validity. An experiment with real users to possess ecological validity must use methods, materials, and settings that approximate the real-life situation that is under study.

Human-Computer Interaction: Human–computer interaction (HCI), also called man–machine interaction (MMI) or computer–human interaction (CHI), is the research field that is focused on the interaction modalities between users and computers (interface). It is a multidisciplinary subject, relating to computer science and psychology.

Location-Aware Computing: Location-aware computing is a technology that uses the location of people and objects to derive contextual information with which to enhance the application behaviour. There are two ways to acquire information about user context: requiring the user to specify it or by monitoring users and computer activity. Sensor technology, such as RFID, could enable mobile devices to extract information from user position automatically.

Mobile Tourist Guide: A mobile tourist guide is a software application with an intuitive interface, that provides users with multimedia information when and where needed during their visit to museums, city centres, parks, and so forth. Such an application runs on PDA-type terminals or on cellular phones, and could be augmented with GPRS (general packet radio service), GPS (global positioning system), and Bluetooth wireless technology. The guide allows tourists to plan routes according to preferences and ambient conditions (weather, timetables, sites of special interest, etc).

Radio Frequency Identification: Radio frequency identification (RFID) is an automatic identification method based on storing and remotely retrieving data using small and cheap devices called RFID tags or transponders. An RFID tag is an object that can be attached to objects, products, or persons to identification using radio waves. Passive tags (with a few centimeter range of sensitivity) require no internal power source, whereas active tags (with more long range of sensitivity, 100 meters) require a power source.

User-Centric Design: User-centric design is a design process that aims at realizing products that meet users’ expectations. The key idea of this design methodology is to start the design strategy taking into account the user’s perspective.

This work was previously published in the Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 657-672, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 3.16
Plagiarism, Instruction, and Blogs

Michael Hanrahan
Bates College, USA

ABSTRACT

This chapter takes as its point of departure the Colby, Bates, and Bowdoin Plagiarism Project (http://ats.bates.edu/cbb), which sought to approach the problem of undergraduate plagiarism as a pedagogical challenge. By revisiting the decision to publish the project’s content by means of a weblog, the article considers the ways in which weblogs provide a reflective tool and medium for engaging plagiarism. It considers weblog practice and use and offers examples that attest to the instructional value of weblogs, especially their ability to foster learning communities and to promote the appropriate use of information and intellectual property.

INTRODUCTION

Alarmist news accounts of student dishonesty and cheating abound. More often than not, such stories describe how universities, colleges, and even high schools have resorted to plagiarism detection services to fight a veritable epidemic of student cheating. The preferred method of combating academic dishonesty, after-the-fact detection, is not the only and is perhaps not the best way to address the problem of student plagiarism. Instead of fighting the lost cause of plagiarism retroactively, technologists and librarians at Colby, Bates, and Bowdoin colleges (CBB) collaborated to develop a program of instruction to educate students about the principles of academic honesty. The resulting plagiarism resource site (http://ats.bates.edu/cbb) includes an introduction to plagiarism, an online tutorial that tests one’s understanding of plagiarism and that provides guidance in the conventions of citation, and a dedicated weblog that publishes links to newsworthy articles, notices, and projects dedicated to plagiarism.

Conceived as a case study, this chapter discusses and evaluates the project’s reliance on a weblog to develop, manage, and publish learning resources dedicated to plagiarism. In the matter of technical choices, the project developers were influenced by their commitment to Open Source Software as well as Creative Commons.
licensing. The former influenced the choice of weblog software, Drupal (http://www.drupal.org), and the latter informed the decision to make all of the project’s learning objects and resources available under an “Attribution-Non-Commercial-Share-Alike” Creative Commons license. These decisions, it turns out, have allowed the project to model the appropriate use of online materials and have retrospectively provided an occasion to reflect on weblogs as an effective medium for engaging plagiarism.

BACKGROUND

Over the past several years, national, regional, local, and campus newspapers across the globe have regularly featured articles on student cheating. While academic dishonesty takes many forms (using a PDA, cell phone, or crib notes during an exam; submitting unoriginal work copied from an existing publication, cut and pasted from an online source, or purchased from a paper mill; or simply peering over a classmate’s shoulder during a quiz), plagiarism has emerged as the most visible form of student cheating. In many ways, the term threatens to subsume all other categories of academic dishonesty. A passing visit to the statistics page at Turnitin’s Web site (plagiarism.org) reinforces this tendency. Turnitin, the world’s leading plagiarism detection service, claims that “A study by The Center for Academic Integrity (CAI) found that almost 80 percent of college students admit to cheating at least once.” Besides generalizing and rounding up the center’s published summary (“On most campuses, over 75 percent of students admit to some cheating”), Turnitin’s claim isolates a common tendency to conflate a number of dishonest “behaviors” with plagiarism. Donald McCabe (personal communication, August 4, 2004) explains that the 75 percent figure published by the CAI “represents about a dozen different behaviors and was obtained in a written survey.” Plagiarism is certainly one form of cheating, but not all cheating is plagiarism.

Reports of plagiarism in the media tend to indulge in hyperbole: it is consistently described as nothing less than an epidemic on campuses. McCabe (1996), who conducted extensive surveys between 1996 and 2003, repeatedly found that the facts do not correspond with “the dramatic upsurge in cheating heralded by the media.” McCabe (2000) has elsewhere observed: “Even though I’ve stated on previous occasions that I don’t believe these increases have been as great as suggested by the media, I must admit I was surprised by the very low levels of self-reported Internet-related cheating I found.” McCabe has subsequently further qualified his view of the problem: “Although plagiarism appears to have remained relatively stable during the past 40 years, . . . it is actually far more prevalent today because many students don’t consider cut-and-paste Internet copying as cheating” (Hansen, 2003, p. 777). More recently, McCabe’s evaluation of his 2002-2003 Survey of U.S. Colleges and Universities identifies an increase in certain kinds of cheating and a continued misunderstanding of plagiarism among undergraduates: “The past few decades have seen a significant rise in the level of cheating on tests and exams. . . . While the data on various forms of cheating on written assignments do not reflect the same trend, this may be due to a change in how students define cheating” (2004, p. 127).

To complicate matters further, statistical estimates of academic dishonesty seem to vary due to contexts (including education level and geography). For example in a recent survey of graduate students enrolled in 32 business programs in the United States and Canada, McCabe, Butterfield, and Treviño (2006) have reported that business students tend to cheat more than other graduate students: “Fifty-six percent of graduate business students, compared to 47 percent of their nonbusiness peers, admitted to engaging in some form of cheating . . . during the past year” (p. 299). The level of self-reported cut-and-paste plagiarism in
this survey, in turn, was “33 percent of the graduate business students . . . compared to 22 percent for nonbusiness students” (p. 300). A recent study conducted by the University of Guelph and co-administered by McCabe and Christensen Hughes (2006) has estimated that 53 percent of Canadian undergraduate students engage “in serious cheating on written work” (Gulli, Kohler & Patriquin, 2007). According to Christensen Hughes, “Serious cheating on written work includes copying a few sentences without footnoting, fabricating or falsifying a bibliography, or turning in a paper that someone else has written” (Cooper, 2007). To help put matters in a global perspective, a recent survey of British higher education conducted by Freshminds.co.uk (with the assistance of the JISC’s Plagiarism Advisory Service and the Center for Academic Integrity) found that “75 percent of respondents have never plagiarized.” This figure in turn approximates what Turnitin representatives have elsewhere estimated: in an interview for the student newspaper at University of California, Santa Barbara, Paul Wedlake, director of sales for iParadigms, the parent company of Turnitin.com, is reported to have claimed that “approximately 30 percent of all students in the United States plagiarize on every written assignment they complete” (Ray, 2001).

Regardless of the figures and statistics, the Internet very much lies at the center of the current fascination with plagiarism. As a result, the fundamentally ethical nature of the offense often gets confused with a technological one. As Patrick Scanlon of the Rochester Institute of Technology has acknowledged: “Plagiarism is not a technological problem—it’s a problem that has to do with ethical behavior and the correct use of sources. And it existed long before the advent of the Internet” (Hansen, 2003, p. 791).

Whether attributed to hype or misperception, plagiarism and the Internet remain entangled in the popular and the academic imaginations. The association is further reinforced by student study habits, especially their research practices. A recent Pew report found that “nearly three-quarters (73 percent) of college students” in the United States claim to “use the Internet more than the library” (Jones, 2002, p. 3). An even greater percentage of students no doubt resorts to the Internet for leisure—to game, surf, IM, and share music files. This reliance on the Internet for study and entertainment has blurred the lines between appropriate and inappropriate cyberpractice and has promoted the intentional as well as unintentional misuse of intellectual and creative property.

The Internet is not the sole source of undergraduate plagiarism. The current manifestation of the problem also can be attributed to non-technological developments, including the increased tendency among students and their parents (at least in the English-speaking world) to perceive higher education as a service industry. That is, the relegation of higher education to a service for which one pays has created a scenario in which students-as-consumers readily expect performance (in the form of good grades) as something to which they are entitled. This sense of entitlement, in turn, overrides concerns about academic honesty. Plagiarism, in this light, emerges as symptomatic of wide-ranging cultural shifts that are not simply or easily reducible to technological shifts and developments. Recent commentary on student plagiarism has provoked observations on this phenomenon. For example, Frank Furedi, professor of sociology at the University of Kent, has observed that “In the ‘customer-client culture’, degrees are seen as something you pay for rather than something you have to learn. It’s the new ethos of university life” (A Quarter of Students Cheating, 2004). This cultural shift and attendant “ethos” may very well lie at the root of the mis-recognition of plagiarism among undergraduates that McCabe has observed (Hansen, 2007, p. 777; McCabe, 2004, 127).

Another significant contributing factor to the rise of plagiarism is an educational culture that resists adapting its instructional methods in the face of advances in technology. This resistance
is forcefully demonstrated by the widespread adoption of plagiarism detection services. In an ostensible attempt to counter technology with technology, schools have settled for a punitive solution to what is a basically an instructional problem, and in doing so have escalated rather than engaged the problem. Turnitin, for example, adds each assignment submitted to its service to its databases. This ethically questionable practice of collecting content has been widely criticized as ignoring the intellectual property rights of students; the issue was raised several years ago by Howard (2001); it surfaced in 2003 at the center of a controversy at McGill University (McGill Student, 2006); more recently Mount Saint Vincent University in Halifax, Nova Scotia, has banned Turnitin for this reason (MSVU bans anti-plagiarism software, 2006); and high school students in suburban Washington, D.C., have protested their school’s subscription to Turnitin on the same grounds (Glod, 2006). In most of these cases, iParadigms has defended its product against this allegation. In a surprising move, however, the company recently took the issue into account when renegotiating its contract with the University of Kansas: “Because Turnitin.com retains student papers, the service has raised intellectual property and copyright issues … Turnitin.com addressed the issue by agreeing to remove papers from the database if requested by the KU Writing Center, which administers the service for KU” (Maines, 2006).

Intellectual property matters aside, the discourse of combating and surveillance that commonly attend the use and promotion of plagiarism detection technology seems ill-suited in an instructional setting. Colleges and universities, after all, have the luxury of privileging learning in their approach to problem solving. Recognizing that after-the-fact detection of plagiarism is a lost cause, faculty, educational technologists, and librarians at Colby, Bates, and Bowdoin jointly developed a plagiarism resource site that attempts to discourage student plagiarism through a program of instruction. The project takes for granted that plagiarism is an inescapable condition of learning. Such a view is by no means unique: Howard (1999, p. xviii), who has published widely on the subject, likens plagiarism to imitation: that is, while trying to find their own voices as writers, inexperienced students invariably adopt and imitate the voices of others and rarely in accordance with the scholarly conventions of attribution. With this view of the problem in mind, instruction would seem to be the desirable as well as the necessary solution to plagiarism. Many educators share this view, and few have been more vocal over the years than librarians, including Burke (2004).

Plagiarism certainly has caught the attention of instructors, librarians, and administrators, but students by-and-large continue to have a vague grasp of it. As Jackson (2006) recently discusses, “there is clearly evidence to support the notion that students, in fact, do not understand plagiarism and lack the necessary skills to avoid it … Many authors agree that students lack understanding of what constitutes plagiarism, how to properly paraphrase, what needs to be cited, and how to cite sources” (p. 420). The many acts of negligence or ignorance that constitute plagiarism also vary in degrees of magnitude: failure to observe accurately the rules for citing sources, for example, is a different order of offense than the inadvertent, unattributed incorporation of another’s language or ideas into a written assignment. These lapses, in turn, are potentially more easily remedied than the conscious, pre-meditated submission of another’s work or ideas as one’s own.

With this range of plagiaristic practices in mind, Howard (1995, pp. 799-800) has usefully identified three categories: outright cheating; non-attribution as a result of unfamiliarity with the conventions of citing sources; and “patchwriting,” or stringing together someone else’s language or ideas without proper attribution. The CBB plagiarism project seeks to promote instruction as the best remedy to help teachers and librarians prevent the last two categories of
plagiarism, which inexperienced students are especially prone to commit. Based on responses to the project’s instructional materials, these goals are being met. For example, Suffolk Community College has used the project’s online tutorial in library workshops on Understanding Plagiarism and Documenting Sources. Students there have found the tutorial helpful, and “they are always particularly interested to learn about the need to cite paraphrases” (Beale 2006). In a recent survey of an online tutorial on plagiarism, Plagiarism: The Crime of Intellectual Kidnapping, created by San Jose State University, Jackson (2006) has produced convincing evidence that “students need more instruction and practice with proper paraphrasing” (p. 426).

To achieve its goal of providing an instructional solution to plagiarism, the project takes full advantage of the Internet and responsible cyberpractice: its developers chose an open source content management system to store, manage, and publish resources; and its resources are freely available not only to be viewed and used via the WWW, but also to be shared, adapted, and re-published under a Creative Commons copyright license.2 The resources include a general overview of academic honesty, an introduction explaining different kinds of plagiarism, an online tutorial for testing one’s understanding of the various practices that constitute plagiarism, and dynamic examples of citations and paraphrasing. The project’s Web site also boasts a dedicated weblog that serves as a clearinghouse on all matters plagiaristic, including news items from around the world and notices on resources, tools, activities, and events concerning plagiarism in higher education. Taking advantage of Web syndication, the project’s weblog makes its content available via RSS feeds. As a result, anyone can import the project’s news updates into individual, departmental, or institutional Web sites or weblogs by means of prepared JavaScripts.3

The CBB Plagiarism Project promotes the responsible use, re-use, and re-purposing of its resources so instructors and librarians can address the problem of plagiarism at the level of local institutional practices, values, and concerns. While plagiarism undoubtedly is a global problem, its solution might best be sought at the local level, where definitions, policies, and expectations vary widely. The decision to publish content by means of a weblog has in retrospect leveraged a technology that has unexpectedly provided a reflective tool and medium for engaging plagiarism. A consideration of weblog practice and use, guided by the concept of plagiarism, provides a framework for understanding the instructional value of weblogs, especially their ability to foster and promote learning communities that discourage plagiarism.

ISSUES, CONTROVERSIES, AND PROBLEMS ASSOCIATED WITH WEBLOGS

Weblogs basically aggregate meta-data: that is, they compile information about information in the form of chronological postings and do not generally publish original content per se. More often than not, weblogs refer and link to other weblogs or Web sites, and the result is a highly interconnected network of communication. The resultant mode of disseminating information has reinforced certain practices that are commonly understood as plagiaristic. Researchers at Hewlett-Packard (HP) Labs have tracked the flow of information in what they call “blogspace” and have identified how ideas, regularly unattributed, spread among blogs (Asaravala 2004). The RSS feeds, moreover, that enable blogs to publish content in various ways are often understood as contributing to plagiarism because they allow unscrupulous users to capture content (specifically textual data) and re-purpose it without attribution. Dishonest practices aside, the HP researchers assert that the dynamic flow of information in blogspace has a generative function: individual weblogs “link
together” to create “a complex structure through which new ideas and discourse can flow.” The HP researchers, Adar, Zhang, Adamic, and Lukose (2004), conceive of the circulation of information among blogs as ultimately creative rather than iterative and original rather than plagiaristic. This interpretation of blogs isolates tensions that have attended the reception of the World Wide Web from its earliest days. Such tensions similarly inform cultural perceptions of our students’ use of the Internet. Their habitual cutting and pasting and sampling and repurposing are commonly dismissed as purposeless, narcissistic self-expression and are censoriously viewed as indicative of their disregard for intellectual and creative property rights and laws. In a recent article, Ellis (2003) productively has situated youth culture’s creative as well as plagiaristic practices in contemporary contexts.

High school and college students operate with the conscious or unconscious understanding (based on a lifetime of practice) that any content available on or accessible via the Web is public property and free. By re-using and re-purposing what they find online, students not only contribute to and reproduce a sub-culture founded on pastiche, but they also develop and acquire the transferable skills that Ellis (2003) suggests will enable those interested to join “the ever-growing ranks of knowledge workers in post-industrial economies.” There are drawbacks as well as benefits to what Ellis envisions as the evolving “new knowledge environment … chunks up human experience into multiple, cross-referenced nuggets dispersed in oceanic cyberspace. Stripped of our distinctively human purposes, the new knowledge environment is what George Trow famously called ‘the context of no context.’” This cutting adrift of knowledge results in its circulation without respect to historical or cultural contexts and creates a number of potential abuses and ethical problems—plagiarism among them. According to Ellis (2003), however, the “new knowledge environment” has some potential benefits that he describes in terms similar to the HP researchers’ description of blogspace: “This environment favors those who can apprehend the interconnectedness of things, create bridges and connections, spark associations and create the éclat of montage. . . . Social network analysis, network topology and other new perspectives are being framed to help us understand the ‘natural’ dynamics of this new environment.”

The dynamics of the blogosphere represent potentially exciting developments in cyber-communication, but they simultaneously revisit many of the criticisms commonly invoked to condemn the WWW. The Web is many things to many people: a commerce tool for business; a recruitment tool for new religions; a protest space for political activism; a play space for dedicated gamers; and so on. Regardless of its intended use or unintended abuse, the WWW has provided interested parties with a readily available means to publish content of all sorts, and its users have responded by taking advantage of its publishing capabilities: according to a recent Pew report, practically half (or 44 percent) of adult users of the Internet in the United States have created and published content (Online Activities and Pursuits, 2004). The value, usefulness, and originality of that content are an endless source of debate, and the popularity of weblogs has provided additional fodder for critics who question the informational and instructional value of the WWW.

Weblogs tend to promote a confessional mode of discourse that celebrates self-referentiality (McDonald 2004). This tendency has fueled the criticism that blogs have ushered in a new era of navel-gazing. The form and content of many personal blogs reinforce this view, but virtual personal diaries do not exhaust the uses and applications of blogs. Adar and Adamic (2005) have suggested that “beyond serving as online diaries, weblogs have evolved into a complex social structure, one which is in some ways ideal for the study of the propagation of information.” Their observation posits an interrelationship of information and its
circulation that previous scholars have variously noted—from McLuhan’s “The Medium is the Message” to Clanchy’s From Memory to Written Record to Brown and Daguid’s “The Social Life of Documents.”

In their approach to the interconnectedness of information and its circulation, Brown and Daguid (1996) have considered the ways in which “developing technologies” have historically “supported social relations in new ways.” A wide range of disciplines and historical examples inform their understanding. Enlisting Anselm Strauss’s notion of “social worlds,” Brown and Daguid (1996) describe a dynamic of group formation that can further the understanding of the culture of weblogs. Following Strauss, Brown and Daguid (1996) observe that “once formed, social worlds continually face disintegration (as dissenting members split off into ‘sub-worlds’).” The blogosphere seems largely populated by such sub-worlds that all too often appear to celebrate a community of one; that is, if one is viewing weblogs as repositories of content rather than as nodes within a network. The flow of information that populates many weblogs, as tracked by the HP researchers, establishes a social matrix that assumes both implicit and explicit communities. Dedicated weblog writers can be roughly divided into two main types: political bloggers and technobloggers. This overly simplistic distinction falls short of capturing the full range of representative blogging sub-worlds (edubloggers, for example), but it conveniently describes two influential communities of bloggers.

Drawing on the theory of the “imagined community” proposed by political scientist Anderson (1991), Brown and Daguid (1996) further consider the ways in which “‘popular’ cultural items, such as journals, novels, pamphlets, lampoons, ballad sheets, and so forth” contributed to the formation of national identity in the American colonies leading up to the Revolution. Citing daily newspapers, in particular, they point out that it was their widespread circulation and not just their content that helped foster Colonial America’s sense of nationhood. The similarities between newspapers and weblogs are instructive. Many observers have noted that blogs have greatly contributed to if not forever changed journalism. McDonald (2004), for example, understands blogging to be “a genuinely positive development in mass communication, and particularly in publishing and journalism.” He attributes the popularity of weblogging to its adoption by “the journalistic establishment.” I would attribute their popularity to their embrace by alternative journalists, especially the proliferation of “warblogging” in the aftermath of 9/11 and the subsequent events leading up to the U.S. invasion of Iraq in 2003.

Weblog pioneer and advocate Dave Winer, moreover, has speculated that newspapers will ultimately be replaced by weblogs as news sources in the not too distant future. This prediction is based in no small part on the publishing ability of weblogs, which has greatly extended the publishing capacity of the WWW. According to Winer, “In a Google search of five keywords or phrases representing the top five news stories of 2007, weblogs will rank higher than The New York Times’ Web site” (Long Bet).

SOLUTIONS AND RECOMMENDATIONS: BUILDING LEARNING COMMUNITIES VIA WEBLOGS

Blogs are powerful and flexible publishing tools: they publish content rapidly and easily; they provide an archive for content that is readily searchable by date, subject, or keyword; and they can also publish their content in a number of ways, including dedicated Web sites as well as RSS feeds that can populate other Web sites, weblogs, aggregators, e-mail clients, and Web browsers. That which has secured their popularity and wide reception (the rapid creation, publication, and circulation of information) also represents
their greatest potential for instruction. Librarians, technologists, and instructors can capitalize on blogs for making available a range of resources and information to targeted users—students, staff, faculty, and colleagues—both on their own as well as on other campuses. They can do so, moreover, with their own content or with content developed entirely by other institutions. This latter ability, importing content from elsewhere, demonstrates how blogs can reinforce the responsible and productive use and circulation of information.

The hallmark features of weblogs (the rapid creation and dissemination of content) are extremely useful for fostering learning communities whose members resort to various methods and media for instruction and information. Certain integral aspects of weblogs further promote their instructional potential. Weblogs have not only made publishing content easier but more social—they open content development up to a group by means of their ability to allow various levels of access to different groups of users; and they invite dialogue between creators and readers of content by permitting the exchange of comments within the blog as well as between blogs. Weblogs are dynamic in a couple of ways: the content posted on them changes as information is added and they allow users to interact by carrying on a dialogue. This dialogic aspect of blogs enables content developers to work towards breaking down the distinction between the creator and the user of content. This feature of blogs participates in the trend already discerned by Pew: that the consumers of Web content are also largely the producers of it.

**FUTURE TRENDS: ENGAGING PLAGIARISM VIA MULTIMEDIA**

The controlled dissolution of boundaries between producers and users of content (or between instructors and students, for that matter) has emerged as a valuable lesson of the CBB project’s use of a weblog. Successful instruction in plagiarism must strive to increase the awareness of the difference between the creation of new and the appropriate use of existing content. The project has sought to promote this awareness in practice by example and in theory by instruction. The content is freely available to be used and re-purposed according to an “Attribution-Non-Commercial-ShareAlike” Creative Commons Deed. The project developers have also sought to create learning objects that help socialize students into the culture of academics, which is founded on what Green (2002, p. 171) has described as the “norm of attribution.” Most teachers take for granted the scholarly conventions used to avoid plagiarism. Recognizing the profound difference that exists between the initiated and the uninitiated, the CBB Plagiarism Project has set out to provide students with guidance and instruction in the standards of academic integrity. In doing so, it strives to facilitate our students’ initiation into the norms and practices of the academic community.

Looking ahead to further development, the project’s next phase will involve creating a more adaptive learning environment for engaging plagiarism. While the weblog provides a valuable means to deliver, create, and respond to content, the text-based nature of that content may reinforce some of the limitations of online tutorials as instructional resources. Jackson (2006, pp. 423-26) has recently considered the effectiveness of plagiarism instruction online. By developing media-rich content about the subject (including audio, video, and animation), the project would create a range of resources that better suit diverse learning styles. In doing so, the project would be more responsive to the needs of its users and would further realize its goal of helping to integrate students into academic cultural practice.
CONCLUSION

An increased use and understanding of media in the curriculum, moreover, may very well allow faculty to harness the creative energies of students in a way that deals with plagiarism in both practical and theoretical terms readily understood by students. Current wisdom on how to avoid plagiarism has emphasized the need to rethink written assignments—for example, essays should be conceived of as ongoing processes consisting of specific, discrete stages or components, all of which are submitted for review, evaluation, and assessment, rather than a single finished product submitted in its entirety only once. In rethinking assignments, instructors may also want to begin to rethink what writing is and to encourage non-traditional forms of writing. I have in mind here the creation of fictional and non-fictional narratives, reports or accounts by means of multimedia—digital video and audio or computer animation and graphics or any combination of these and other media. Just as the weblog has emerged as a reflective tool for considering plagiarism, a media-rich learning environment would allow students to begin to understand plagiarism in new and perhaps more compelling ways. In a recent essay on plagiarism, the novelist Jonathan Lethem (2007) describes what it is like to be cut adrift in our contemporary media environment:

The world is a home littered with pop-culture products and their emblems. I also came of age swamped by parodies that stood for originals yet mysterious to me … I’m not alone in having been born backward into an incoherent realm of texts, products, and images, the commercial and cultural environment with which we’ve both supplemented and blotted out our natural world. I can no more claim it as “mine” than the sidewalks and forests of the world, yet I do dwell in it, and for me to stand a chance as either artist or citizen, I’d probably better be permitted to name it.

In the academy, students are encouraged to name and when appropriate cite their sources, influences, and inspirations. However, finding themselves, like Lethem, in a world already created and populated with signs, they need to learn how to negotiate the conventions and practices of that world and to decode its constituent signs. Educators should begin to make use of the multitude of media that figures our manifold experiences of the world. The energy and creativity generated by such a diversely constituted learning environment would permit powerful models for rethinking our engagement of plagiarism.

REFERENCES


*Center for Academic Integrity.* Retrieved October 2, 2006 from http://www.academicintegrity.org/cai_research.asp


This work was previously published in *Student Plagiarism in an Online World: Problems and Solutions*, edited by T. Roberts, pp. 183-193, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).


**ENDNOTES**

1 Original project members included Judy Montgomery and Sue O’Dell, Bowdoin College; Zach Chandler and Marilyn Pukkila, Colby College; and Thomas Hayward, Bates College. Jim Hart at Bates College served as a technical consultant from the project’s inception and generously provided extensive support by administering the Linux server that hosted the project’s weblog and resources.

2 The site is driven by Drupal, a PHP-MySQL-based open-source content management system, which is freely available to download at http://www.drupal.org. For further details on Creative Commons, see http://creativecommons.org.

Chapter 3.17
Twin Wiki Wonders? Wikipedia and Wikibooks as Powerful Tools for Online Collaborative Writing

Meng-Fen Grace Lin
University of Houston, USA

Curtis J. Bonk
Indiana University, USA

Suthiporn Sajjapanroj
Indiana University, USA

ABSTRACT
Web 2.0 technologies empower individuals to contribute thoughts and ideas rather than passively survey online content and resources. Such participatory environments foster opportunities for community building and knowledge sharing, while encouraging the creation of artifacts beyond what any single person could accomplish alone. In this chapter, we investigate the emergence and growth of two of such environments: the highly popular Wikipedia site and its sister project, Wikibooks. Wikipedia has grown out of trends for free and open access to Web tools and resources. While Wikipedians edit, contribute, and monitor distinct pieces of information or pages of documents, Wikibookians must focus on larger chunks of knowledge, including book modules or chapters as well as entire books. Several key differences between these two types of wiki environments are explored. In addition, surveys and interviews, conducted with Wikibookians, shed light on their challenges, frustrations, and successes.

INTRODUCTION
Thomas Friedman, in his 2005 book, The World is Flat, talks about 10 forces that have flattened the world in terms of economic globalization. The word “flat” acts as a metaphor to symbolize the “leveled” playing field on a global scale. In Friedman’s (2005) view, when the playing field is leveled, everyone can take part. And he means...
everyone! Talking from a business perspective, Friedman charts the progress of globalization from what he describes as 1.0 to 3.0. Globalization 1.0 focused on country to country relationships, such as treaties and trade. In Globalization 2.0, such relationships moved down to a company-to-company level. We are now entering an age of Globalization 3.0, where the rise of the individual comes into focus. It is in this third phase of globalization, wherein individuals obtain new powers and freedoms to participate socially, economically, politically, and educationally with others around the world. *Time Magazine* recognized this trend, and in 2006 named “You” as the person of the year (*Time Magazine, 2006/2007*). In effect, the Year 2006 signified the trend toward empowering technology users with what was called Web 2.0 technologies, which allowed individuals to generate ideas online instead of just reading and browsing through someone else’s Web pages (*Grossman, 2006/2007a, 2006/2007b*).

Lessig (2005) called this phenomenon the “read-write Web,” to contrast the read-only Web, where users were merely passive consumers of information. Most fundamentally, the read-write Web dramatically enhances the power of individuals, and fosters a participatory culture of building, tinkering, learning, and sharing (*Brown, 2006*). Typically included in Web 2.0 technology lists are wikis, podcasts, blogs, online photo albums, and virtual worlds such as Second Life. In particular, wikis offer an innovative mechanism of computer-supported collaboration. This highly unique social phenomenon utilizes armies of volunteers who work without pay, recognition, or much acclaim to continually contribute, update, or edit resources online (*Lin, 2006*).

In this chapter, we primarily offer insights into the recent emergence of two wiki developments; namely, Wikipedia and Wikibooks. First, we detail what a wiki is. Subsequently, we briefly review existing wiki-related research, and then document the emergence and growth of what is the most popular wiki today; Wikipedia. After presenting our research findings in Wikibooks, we compare and contrast some of the underlying principles and features of Wikipedia to those found in Wikibooks.

**BACKGROUND**

Brandon Hall (2006) defines a wiki as “a collection of Web pages that can be easily viewed and modified by anyone, providing a means for sharing and collaboration.” These are open-ended, generative, and unstructured environments (*Honegger, 2005; Leuf & Cunningham, 2001; Lio, Fraboni, Leo, 2005*). Pioneered by Ward Cunningham in 1995, wikis are online spaces for recording information, sharing knowledge, typically in collaboration with others. Each modification is recorded as the history of a document. The history page records the time of change, the person who made the change, and the changes that were made. Such a mechanism not only permits page retraction by anyone, it also behaves as a podium for reputation management. In addition, the history page permits open examinations of each revision, allowing each version to be compared and contrasted by anyone.

Many universities have picked up the wiki fever and started using its functions for information sharing. For example, Stanford has an institutionalized wiki wherein students can contribute information on places to eat, workout, study, socialize, and so forth. (*Campus Technology, 2006b*). As this Web site indicates, there is now a wave of student-contributed wiki resources. Similarly, MIT has created the Center for Collective Intelligence, where people from around the planet could come and solve huge scientific, social, and business problems (*Campus Technology, 2006a*). The underlying belief of these wiki projects indicates that, collectively, the human race can act more powerfully than it can at an individual level. As a prime example of this principle, in early February, 2007, Penguin books announced
“A Million Penguins,” a Web site “where people from around the planet will help write a novel” (Reuters, 2007). Similarly, in the academic world, faculty at MIT and Wharton School of Business invited thousands of individuals or potential authors to help write a collaborative online textbook at the site “We are Smarter than Me” (Campus Technology, 2007).

There is scant academic research on wiki technology and resources even though we are in the midst of a proliferation of wiki-based projects. In particular, the research on Wikibooks, which emerged with much enthusiasm in 2003 (Wikibooks, 2007a), is particularly thin. Some Wikipedia-related research has explored the accuracy of the content (Lih, 2004; Rosenzweig, 2006), historical flow visualizations of contributions (Viegas, Wattenburg, & Dave, 2004), the formality of the language used (Emigh & Herrring, 2005), hackers and trolls (Schachaf & Hara, 2006), contributions within collaborative authoring environments (Korfiatis, Poulous, & Bokos, 2006), differences between experts and novices (Bryant, Forte, & Bruckman, 2005), and cultural differences in Wikipedia contributions and editing behaviors (Pfeil, Zaphiris, & Ang, 2006). Such studies are, of course, just the starting point for a long line of research on wikis.

A BRIEF HISTORY OF THE BRIEF HISTORY OF WIKIPEDIA

Wikipedia is a free online encyclopedia to which anyone can contribute. Created in January 2001, Wikipedia started out as a side experiment of a failing project called Nupedia.com, created by Larry Sanger and Jimmy Wales, that also aimed to create a free online encyclopedia. However, Nupedia.com employed a traditional prepublishing review process that demanded an elaborate and often long procedure of coordination among submission, reviewing, and negotiation. Few articles were published as a result of this arduous process. Wikipedia was created as an experiment to bypass this prepublishing review process and empower the postpublication, real-time, peer-review procedure among volunteer collaborators (Ciffolilli, 2003; Voss, 2005).

Today, Wikipedia is perhaps the largest instance of a wiki. With easy to explore and open technology at its core, Wikipedia quickly became an online volunteer community with its devotion to the creation of a free encyclopedia where the division of labor is facilitated by information technology. That volunteer community now includes more than 50,000 individuals who have made at least 10 contributions to Wikipedia (Rosenzweig, 2006).

According to Alexa, a Web ranking service, as of February 7th, 2007, Wikipedia had become the 12th most visited Web site across all languages (alexa.com, 2007). During the past 6 years, Wikipedia has enjoyed tremendous, if not exponential growth in terms of registered users, number of articles, and number of languages. Not surprisingly, Wikipedia quickly overshadowed Nupedia.com and caused Nupedia.com to close in 2002. As of February 2007, the combined Wikipedia site included articles in 250 languages, 6.4 million articles, and almost 3.5 million contributors (Wikipedia, 2007d). The English language version of Wikipedia continues to be the largest site at more than 1,624,000 articles, while some other language versions remain quite small (Wikipedia, 2007b, Wikipedia, 2007c). There are more than 300,000 articles in German, French, Polish, and Japanese languages, with the German Wikipedia site now pushing over 500,000. This is an impressive amount of work to accomplish since inception 6 years ago. It is also quite obvious that a vast number of willing volunteers awaited just such an opportunity to contribute their time and knowledge.

Wikipedia’s exceptional growth was perhaps best captured and described by Voss (2005), who conducted one of the first quantitative analyses of the structure of Wikipedia. From the elaborated
numbers, charts, and mathematic formulas, Voss presented a comprehensive picture of the overall Wikipedia community. As would be expected, Wikipedia experienced a period of linear growth, while the more spiked or exponential growth that is often alluded to in the media did not occur until about 15 months after inception in April 2002. Since that time, the number of new articles in Wikipedia has increased at various speeds among different languages. Even though anyone with Internet access on this planet is invited to contribute to Wikipedia, according to the Voss study, the average number of authors per article remains four to five, a number that is perhaps much lower than most people would assume. Even more interesting is that about half (48%) of the articles had less than five authors and about one-third (28%) of the articles in German Wikipedia had only been edited by one author. In addition, one-third of the Wikipedia authors had only contributed one article, and just one in five Wikipedians had been involved in more than 16 articles. Granted that Voss excluded anonymous edits in his calculation, these numbers indicate that not everyone will contribute, or at least continue to contribute, just because they can.

In summary, then, the Wikipedia community has the following characteristics. It is a volunteer-based project to which everyone can contribute. Wiki technology facilitates certain divisions of labor, enables the ease of contributing and participation, and records all change histories for easy examination and recovery. It is a rapidly growing community where authority is gained through individuals’ active participation. Its social structure does not necessarily recognize people based on their expertise, even though it does employ a promotion mechanism for various roles in the community. Positioning itself as an encyclopedia, Wikipedia inherits the perception of being a reliable reference source. However, its credibility continues to be questioned.

**WIKIBOOKS AND OTHER WIKI SITES**

Sister projects to Wikipedia coordinated by the Wikimedia Foundation include Wikibooks, mentioned previously, as well as Wikispecies, Wikiquotes, Wikinews, Wikiversity, Wiktionary, Wikisource, Commons, and Meta-Wiki. Wiktionary, for instance, is a free, community created online dictionary and thesaurus with more than 300,000 entries in nearly 400 different languages (Wiktionary, 2007). Wikiquotes contains an index of thousands of quotations from famous people from Laozi, William Shakespeare, Aristotle, Helen Keller, and Martin Luther King, as well as from literacy works, television shows and movies, news pages, and other creative efforts (Wikiquotes, 2007). Wikinews, like other online news services, contains current events making headlines. But as folks like Rosenzweig (2006) point out, it often can break leading stories before the other news services since its reporters cover the planet.

In our own work, Wikibooks has been the primary focus. The Wikibooks Web site, originally named the Wikimedia Free Textbook Project and Wikimedia Textbooks, was initiated on July 10, 2003 (Wikibooks, 2007a). It was the brainstorm of Karl Wick, who was searching for a place to host his books on organic chemistry and physics as a means to bring such educational resources to the people of this planet while addressing the high costs and other limitations of existing books and materials on these topics. By the end of that month, there were 123 modules or chapters posted and 47 registered users at that site. Two months later, there were 530 modules and 150 registered users. Since it began, controversies and criticisms have been posed regarding the scope of the Wikibook project, copyright and licensing issues, the quality of the books posted, and the large number of unfinished book projects.
The coordinator of one or more Wikibook projects is called a Wikibookian. Given that in less than 4 years Wikibooks has become a collection of over 1,000 free online book projects containing more than 20,000 modules or chapters (Wikibooks, 2007b), there are many active Wikibookians. The Wikibooks Web site contains collaboratively written textbooks and nonfiction books, study guides, and booklets are socially negotiated documents. There are both books for adult learners as well as the more recent emergence of junior books for younger learners, ages 8 to 11 (Wikibooks, 2007c) (see Figure 1). The junior Wikibooks site is currently creating full-color booklets on topics such as bugs, big cats, the kings and queens of England, and dinosaurs.

Such environments offer hope for someday providing access to educational books, study guides, and other documents to every connected learner and in any language. Nevertheless, there are myriad issues, questions, and problems related to online collaboratively authored books, such as...
Twin Wiki Wonders?

as those found at the Wikibooks site, that need to be better understood. Figure 2 depicts the key aspects of the Wikibook environment.

Studying Wikibookians

During a 4-month period in early 2006, we observed how Wikibookians communicated and shared ideas with each other. Based on our observations and literature review, we designed a close-ended survey that encompassed basic demographic data, as well as a set of questions geared toward issues involved in the process of creating a Wikibook from a sociocultural point of view (Bruns & Humphreys, 2005). Through communication channels frequented by active members, we obtained a user list providing around 45,000 Wikibookian's names who had previously registered an account at the Wikibook Web site. Using this list, in the spring of 2006, we randomly sent survey invitations to around 1,500 Wikibookians whose status was shown as active via the contacting function found at the Wikibook Web site. Using this list, in the spring of 2006, we randomly sent survey invitations to around 1,500 Wikibookians whose status was shown as active via the contacting function found at the Wikibook Web site. We used an online survey tool, SurveyShare, to distribute the survey and collect the data. We received 80 responses to our survey. Ninety percent of the survey respondents had been active in Wikibooks for more than 6 months, 72% for more than one year, and nearly one-third for 2 years or more. In addition, 82% were under age 35, 97% were men, and more than half did not have a four-year college degree (Sajjapanroj, Bonk, Lee, & Lin, 2007).

After tallying the surveys, we randomly selected 15 people from the survey respondents for follow-up e-mail interviews. The e-mails were sent in July 2006, and eight Wikibookian participants joined this part of the study. Participants were asked to answer a 12-question e-mail interview related to how they had started in Wikibooks, previous experiences in such an online collaboration process, the future of Wikibooks, advantages and disadvantages of this environment, whether an online book could ever be complete, and suggestions for improvement.

FINDINGS

The survey and interview data helped us understand the challenges, frustrations, and successes of Wikibookians. From the 8 participants and 12 e-mail interview questions, five key themes emerged. These themes were as follows:

1. The expression of authority in Wikibooks
2. The nature of collaboration process in Wikibooks
3. Completeness and success of one’s Wikibook project
4. Advice to Wikibook wannabes
5. The future of Wikibooks

The following sections describe the themes and present the different views participants offered. Based on these findings, as well as existing wiki literature, we offer a comparison of the Wikipedia and Wikibook environments.

The Expression of Authority in Wikibooks

Authority

Working on a wiki as an editable Web platform could be rewarding and challenging at the same time. The number one critique of Wikipedia has been its authoritative value. It is not surprising that such questions and concerns about authority also appeared in our Wikibook findings. Some participants seemed to disregard the traditional view of an “expert” or authority:

...people can work together on a wiki and come up with a result that is better than something written by one or a couple of “experts.” ...There is not one
person in charge who can make the hard decisions that everyone will respect. (Participant 1)

Nevertheless, some seemed to also demand the skill of “experts” in order to create a good quality book:

I expected it to be not very effective, as it’s hard to let people work together at all, especially on this large amount, and even more when it’s spread over multiple pages with a special subject. You should need at least two specialists, as both have to understand the subject very good. That appeared to be hard indeed. (Participant 6)

Wikibookian views about community or multiple contributor authority over the knowledge of “experts” revealed the nontraditional belief that each one of us possesses our unique knowledge that could contribute to the creation of a Wikibook. On the other hand, there seemed to be a demand for such knowledge to be able to qualify at an acceptable level as an expert.

Control

If looking at the issue of authority at a deeper level, we further discovered the concerns regarding control over contribution. Who has the right to contribute? Who can decide who has the right to contribute? Who can decide whose contribution is superior to others? In other words, is it true that in the world of Wiki, everyone stands equally? Certainly some did not believe so:

The biggest problem I see are the control guilds that have sprung up. They can be the Christian groups, or the anti-metaphysicists that go around labeling everything pseudoscience because they lack the ability to see the difference between the two. (Participant 2)

Most participants addressed the issue about control over contribution from the perspective of being able to monitor changes. For instance, Participant 7 wrote to us about the importance of being able to “track the contributions and changes of each contributor.” Participant 3 elaborated further:

I think the fact that Administrators monitor the content and make sure that if it is inaccurate or copyright protected that it cannot be posted is a very important aspect of WikiBooks/Wikipedia that should be emulated by others. (Participant 3)

One other aspect of control over contributions surfaced as well. As shown in the following quote, Participant 6 expressed the desire to create a temporarily safe environment during the creation of a Wikibook:

Make a special area where one set group of people can take over a book for a time, for example, to enable one class or one group of professors to develop materials in a protected environment where, at least for a time, they have the final authority of whatever happens in that area. (Participant 6)

As indicated, Participant 6 believed that having authority even for a period of time was deemed important. Participant 4 shared that “the biggest disadvantage is a sacrifice in control over the direction of the book,” which seemed to indirectly echo Participant 6’s point. Here, the control issue seems to center on setting the overall goal of the book and the initial creation of the first draft.

As demonstrated, even though most participants did not directly bring up the question about control, through interview excerpts, we observed how multiple dimensions of authority were realized among Wikibookians. Our survey data indicated that an overwhelming majority of the surveyed Wikibookians believed that Wikibooks promote individual accountability in terms of one’s contributions to an online book project (15% strongly agree, 54% agree). However, still
31% of them disagreed that individuals could be held accountable for the accuracy of contents contributed. It is not surprising to see that Wikibookians wrestled with these questions, which leads to the next section about how Wikibookians collaborate.

The Nature of Collaboration Process in Wikibooks

Coordination

Working in a Wikibook project requires coordination, communication, an understanding of the collaborative writing process, and a great deal of time management. In particular, Participant 4 shared many of his perspectives about the inner workings of a Wikibookian and the process of Wikibook creation:

Development has been slower than expected. I expect to be able to finish the book...in about two years.

My experience has been that Wikimedia projects promote much more formal rules of etiquette, but a much less structured writing process.

Maintaining a wiki is much more a challenge for social issues than for technological issues. Above all, a collaborative writing community must have a common vision, specifically a vision to create a written work.

Interestingly, he suggested that software be created that makes it easy to post a positive contribution and more difficult to post something negative.

Communication

Clearly, issues about communication stood out in many of the Wikibookian observations and suggestions. Some of the comments included:

Coordination is key. So [is] communication between editors, contributors, etc. is important. It must also be easy to use, fast and one must quickly be able to see what was changed, who changed the document and if necessary, change it back. (Participant 8)

Collaboration

The concept of a wiki assumes a collaborative process. While many participants clearly showed their agreement, two indicated their disappointment when no other took interest in their book:

I also hoped that collaborators would take an interest in the book as it developed. Although there have been isolated contributions, no one else has taken an interest in making a real investment in the book. I still hope this will eventually happen. (Participant 4)

My expectation was to help create a free, collaboratively written textbook. At this point I have done almost all of the work on the book. Admittedly, word about the book hasn’t gotten out, but hopefully once it does others will begin to contribute. Until that happens, my expectations will not, technically, have been met. (Participant 7)

Resolving Disagreements

Even though some seemed to be lone writers, disagreements are bound to happen when working with strangers from around the world. We asked the Wikibookians about how they resolved differences. Again, communication provided the bridge to reconcile different opinions, including mechanisms such as the talk page, a mediator, private message, or revert the changes. However, Wikibookians differed in how they would approach or resolve the differences. As the following quotes indicate, some would revert the changes, some would discuss them first, and still others
did not know since they had yet to experience such situations:

So revert it :) It’s a Wiki, so everybody can edit it. When s.o.’s edit doesn’t apply to my standards, I can revert it. And that person can revert me too. When we both find it important, we can start talking through the talkpage, and in the end, maybe get a mediator, or ask the community what they think about it. (Participant 1)

This happens frequently. I usually either talk to them via the “Talk” pages or send them a private message regarding the change(s) if they are not an anonymous user. I leave or qualify their changes if they are correct, but if not, I may revert them back to their original form. (Participant 3)

Perhaps such quotes reflect the degree of expertise of the Wikibookian. Byrant, Forte, and Bruckman, (2005), for example, found that expert Wikibookians have a policy not to revert changes or additions in Wikipedia if at all possible. They tend to post their concerns to the talk page first.

From the nature of the collaboration process, we see the complexity of not only the communication process among authors, but a philosophical question about the definition of a finished Wikibook emerged. As an emerging phenomena, Wikibooks draw interest and attention, especially for their education and self-learning value.

**Completeness, Success, and Happiness Regarding One’s Wikibook Project**

**Wikibook Completion**

Given the comments and concerns about the quality of Wikibooks and length of time to complete them, it was important to ask Wikibook contributors about Wikibook completion and overall successes experienced. Using all the methods mentioned here, will the revision process ever be completed? In other words, will a Wikibook ever be finished? We raised this question to all of our participants and had rather interesting findings. In our survey responses, nearly 60% of Wikibookians indicated that a Wikibook could be completed. However, as the following quotes exemplify, respondents to our interview questions were quite varied in their opinions regarding online book completion:

It will probably depend on the subject. When you will try to explain a limited subject, it can theoretically be complete. However, it will be hard to have it finally completed indeed. (Participant 1)

No wiki is ever complete, because it is ever evolving. That’s one of the best things about wikis. I personally think that paper is dead and in many ways the ideas contained within them too. I want my ideas and thoughts evolved and allowing others to improve them makes the work alive. (Participant 2)

Wikibooks can eventually become “complete,” when the authors decide that there is no more to write...so a book is substantially complete once the last “proper” author leaves....It is by dedicated writers that books are written, not casual one-sentence contributors or people who drop in for an afternoon. (Participant 4)

As demonstrated, some Wikibookians thought that completion of the book would naturally be indicated by participant interest in the topic or the scope of the project. One Wikibookian explained that “I think a Wikibook becomes complete when the participants loose interest in the topic.” On the other hand, another Wikibookian suggested that a Wikibook could be complete if all the relevant information on a topic was included or known such as the “Iran-Contra scandal of the early to mid 1980s.” Overall, however, there was some
agreement that a Wikibook is not as a product but a process, because a Wikibook is always evolving and “allowing others to improve them makes the work alive.” From this notion, a more refined question surfaced: who decides whether a Wikibook is finished? Can we say a Wikibook is finished when nobody is contributing to it anymore?

**Wikibook Success and Happiness**

According to our online survey, most Wiki-bookians felt that their Wikibook was a success (see Figure 3). We did not, however, ask what determines that success. This issue relates back to the previous one regarding completion; if Wiki-bookians focus on the process and not the final product, then success is an extremely fleeting concept. At the very least, success is then focused on social interaction and collaboration processes, instead of simply on the quality and use of the final document.

Some Wiki-bookians seem truly inspired by the Wiki-book process and working in any type of wiki environments. For example, Participant 1 simply stated, “Go rockin’ on!” when asked for additional comments, while Participant 2 noted, “I love wikis they’re truly the closest example of the purest form democracy.”

**Advice to Wiki-book Wannabes**

**Technology Skills and Realistic Goals**

Obviously, a Wiki-bookian needs to have the necessary technology skills to operate within a wiki environment. Beyond technological savvy, understanding the magnitude of creating a book and setting realistic goals could help start one on the right foot:

*A little bit is better than nothing, it can always be improved later. Often times there is no book for a topic because everyone thinks “Oh, someone else will make it.”* (Participant 6)

*Get help. Don’t try to do it on your own, it’s a too big amount of work and you will definitely loose the overview.* (Participant 1)

---

![Figure 3. Wiki-bookians perceptions of Wiki-book success](image-url)
Wikibooks are Demanding

As the following quote from Participant 5 indicates, experienced Wikibookians also warned us about the difficulty of chapter organization:

*I'd suggest getting several coauthors from the get go and deciding on a template for the book chapters so it is uniform from the beginning. It's bound to change over time, but you may as well start with a plan. Also, I'd mention that, for the most part, they shouldn't expect many people to contribute to the book. Writing a book is challenging; most people will never attempt it.* (Participant 5)

In addition to getting the flow and direction of the book, Participant 4 advised us the importance of understanding the book-writing process. In this opinion, joining an existing book might be the best way to become involved in Wikibooks:

*I would first offer a warning that writing can be very difficult. Writing a textbook is an involved task, demanding the full scope of your expertise, and requiring a substantial investment before any payoff can be realized. Yet it's extremely rewarding......I would encourage users to consider working on an existing book rather than starting a new one. Starting off with a more manageable goal will let a user test the water without making a big commitment.* (Participant 4)

Wikibooks are Different from Wikipedia

Most Wikibookians seem to have initial wiki backgrounds from editing or contributing to Wikipedia. They may have shifted to Wikibooks on the advice of others that their textbook-related ideas belonged in Wikibooks not in Wikipedia. Still, their prior history in Wikipedia shaped their expectations. Among all the suggestions, Participant 5 brought up the idea that involvement with a Wikibook is different from working on a piece of an article in Wikipedia:

*I think authors have to make concepts before they write a book and think about it, because it's not only an article like in Wikipedia.* (Participant 5)

As a result of this difference, in order to work in a Wikibook, more communication, organization, and compromise is required. Participant 3 shared with us his view about how Wikibooks are different from other collaborative productions:

*I expected WIKIBOOKS to be more in-depth about specific subject areas and to provide more details than Wikipedia articles. In the case of video games, for example, I expected WikiBooks to provide background stories and walkthroughs.* (Participant 3)

Future studies might explore the need for additional planning processing and tools within Wikibooks. Would outlining and planning tools, for instance, enhance the final product and the percentage of Wikibooks that are, in fact, completed? These and other differences between Wikipedia and Wikibooks are explored later on in Table 2.

The Future of Wikibooks

We were interested in how these experienced Wikibookians viewed the future of their Wikibook work. Based on their experience, did Wikibooks even have a future? And what tools and features might enhance it? As Participant 6 rightfully pointed out, the spirit of a wiki aims to “make the world a juster place by putting the tools of education into the hands of every person.”

Future Growth and Success of Wikibooks

Some seemed concerned about the lack of reader interests at this early stage of Wikibooks:
Twin Wiki Wonders?

It’s difficult to tell at this stage. Wikibooks is faced with a sort of “chicken and egg” problem -- it needs readers to attract writers and writers to attract readers. The project will recruit more and more talented writers as it picks up steam, and will have better and better books in its library. ... But Wikibooks may never take off at all. ... (Participant 4)

The so-called “chicken and egg” issue was shared by Participant 1, who believed Wikibook will be either successful or die. As he stated, “There is no ‘some people will use it.’ Because when you want your book to become used, it has to be used by a large amount of people, not by a few.” He also felt that a single dedicated person could make a huge difference in determining the ultimate success or failure of Wikibooks. “Now that is some incentive to participate!”

Then the next logical question arises. Why do few people use the books at the Wikibook Web site? Is it because “there has not been an example yet of a book that goes all the way through and is put into use” or perhaps it is because “People have not seen it work enough for them to believe in it, and there are not mechanisms in place that convince educators of the value of Wikibooks as a teaching or learning tool?”

Academic Acceptance

In their responses to our online survey questions, Wikibookians seemed to strongly believe in Wikibooks as an online library (64%), a learning tool (40%), and a supplement to classroom or training resources (36%). In addition, it was a place for communities of writers (60%) and learners (34%) to form. However, making Wikibooks a teaching and learning tool and enabling the educational community to embrace it proved a difficult task. Academics might be the largest potential users of Wikibooks; they might hold the key to Wikibooks’ success. The following quotes indicate the shared concerns and hopes of Participant 6 and 7:

I don’t think the concept will catch on except among small niches until there are some very well-developed textbooks available on the site. If academics turn to using it regularly for free introductory textbooks, Wikibooks could explode with interest. But we aren’t there yet. Maybe in the next 5 years or so. (Participant 7)

This is a very valuable project with very important goals, and any educator who is able to contribute a bit or to draw his or her students or colleagues into the process will be doing groundbreaking work in education which is worthy of praise. Things can only continue to go up from here. (Participant 6)

FUTURE TRENDS

Each of the previously mentioned themes provides interesting data for further wiki-related research and development; especially that related to Wikibooks. Table 1 summarizes our major findings according to the five themes.

COMPARING WIKIPEDIA AND WIKIBOOKS

As indicated in the literature review, there are many distinct differences between Wikipedia and Wikibooks. As shown in Table 2, they differ in terms of the types of communities that they foster, the resources created, size of user community, technology supports required, the goals and needs of both experts and novices, and many other variables.

FUTURE RESEARCH DIRECTIONS

After their birth in 1995, wikis proliferated quickly into the mainstream of human life. From users’
**Table 1. Key themes from interviews with Wikibookians**

<table>
<thead>
<tr>
<th>Theme 1: The Expression of Authority in Wikibooks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Authority:</strong> Each one of us possesses unique knowledge that could contribute to the creation of a Wikibook.</td>
<td>“not one person in charge”  &lt;br&gt;“need at least two specialists, as both have understand the subject very good”</td>
</tr>
<tr>
<td><strong>Control:</strong> Some liked to be able to monitor changes and most believed in personal accountability.</td>
<td>“When a lone writer is trying to open people’s minds to the reality of things and closed minded indoctrinated individuals that in all reality could be a 13 yr old who has been given admin powers can be judge and jury over your work..”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theme 2: Collaboration Process in Wikibooks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coordination:</strong> Writing a Wikibook together might be more of a social issue than a technical one.</td>
<td>“Above all, a collaborative writing community must have a common vision”</td>
</tr>
<tr>
<td><strong>Communication:</strong> Communicating among writers proved essential.</td>
<td>“…requires communication in order to get the project off the ground and into something tangible…”</td>
</tr>
<tr>
<td><strong>Collaboration:</strong> Finding someone else who is also interested in your book might not be so easy.</td>
<td>“…there have been isolated contributions, no one else has taken an interest in making a real investment in the book…”</td>
</tr>
<tr>
<td><strong>Resolving Disagreements:</strong> One would revert the changes, some would discuss them first, and still others did not know since they had yet to experience such situations.</td>
<td>“I usually either talk to them via the ‘Talk’ pages or send them a private message regarding the change(s) if they are not an anonymous user.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theme 3: Completion, Success, and Happiness Regarding One’s Wikibooks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completion:</strong> Agreement seemed to be that a Wikibook is not a product but a process because a Wikibook is always evolving.</td>
<td>“No wiki is ever complete, because it is ever evolving. “  &lt;br&gt;“Wikibooks can eventually become ‘complete,’ when the authors decide that there is no more to write”</td>
</tr>
<tr>
<td><strong>Success and Happiness:</strong> Success is then focused on social interaction and collaboration processes.</td>
<td>“Go rockin’ on!”  &lt;br&gt;“I love wikis they’re truly the closest example of the purest form democracy.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theme 4: Advice to Wikibook Wannabes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technical Skills and Realistic Goals:</strong> Be realistic regarding your skill level and writing goals.</td>
<td>“Don’t try to do it on your own, it’s a too big amount of work and you will definitely loose the overview.”</td>
</tr>
<tr>
<td><strong>Wikibooks are Demanding:</strong> Writing a book chapter is a difficult task.</td>
<td>“Writing a textbook is an involved task, demanding the full scope of your expertise, and requiring a substantial investment before any payoff can be realized.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theme 5: The Future of Wikibooks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wikibooks are Different from Wikipedia.</strong></td>
<td>See Table 2 for comparison.</td>
</tr>
<tr>
<td><strong>Growth and Success:</strong> Lack of readership at early stage.</td>
<td>“it needs readers to attract writers and writers to attract readers.”</td>
</tr>
<tr>
<td><strong>Academic Acceptance:</strong> Academics might be the largest potential users of Wikibooks; they might hold the key to Wikibooks’ success.</td>
<td>“…educator who is able to contribute a bit or to draw his or her students or colleagues into the process will be doing groundbreaking work in education which is worthy of praise.”</td>
</tr>
</tbody>
</table>
## Twin Wiki Wonders?

### Table 2. Comparison of Wikipedia and Wikibook environments

<table>
<thead>
<tr>
<th>Issue or Characteristic</th>
<th>Wikipedia</th>
<th>Wikibook</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Community</td>
<td>Wikipedia is a community of practice for millions of people. There are myriad subcommunities within it for different languages, topics, resources, and tools.</td>
<td>Wikibooks has communities of practice for each book project. There is also an overall community of Wikibookians at the staff lounge within Wikibooks.</td>
</tr>
<tr>
<td>3. Resources Created</td>
<td>Wikipedia is an information resource for people to look up. It is comprised of many linked, yet individual pieces.</td>
<td>Wikibooks creates usable texts, guidebooks, and reference materials. The final product should be coherent.</td>
</tr>
<tr>
<td>4. Historical Statistics (as of February 9, 2007):</td>
<td>7,483,939 pages 6.4 million articles 1,629,257 articles in English 250 languages 110,836,256 edits 14.81 edits per page 700,001 media files 3,511,411 registered users 1,111 system administrators (Wikipedia, 2007d).</td>
<td>67,399 pages 23,790 modules or chapters Over 1,000 books, the largest category in English 120 languages 759,033 page edits 11.26 edits per page 50,582 registered users, 33 system administrators (Wikibooks, 2007b).</td>
</tr>
<tr>
<td>5. Views of Contributors</td>
<td>“One could think of the Wikipedians as a global ant colony, except that there is no queen, and users choose their own roles.” Some Wikipedians welcome newcomers, while others determine awards, upload images, clean article grammar, work on article histories, revert vandalism, and so forth. (Wikipedia, 2007e).</td>
<td>A Wikibookian is someone who coordinates or contributes to a Wikibook project.</td>
</tr>
<tr>
<td>6. Goals and Focus</td>
<td>Wikipedians are concerned with the quality of Wikipedia. They maintain or monitor a set of “watch” pages (Bryant, Forte, &amp; Bruckman, 2006).</td>
<td>Wikibookians are concerned with the quality of the particular books and study guides that they are preparing.</td>
</tr>
<tr>
<td>7. Participation or Contribution Criteria</td>
<td>Enter Wikipedia site with generalized knowledge but add specific pieces of information.</td>
<td>Enter book project with specialized knowledge.</td>
</tr>
</tbody>
</table>
Table 2. Comparison of Wikipedia and Wikibook environments (continued)

<table>
<thead>
<tr>
<th>Issue or Characteristic</th>
<th>Wikipedia</th>
<th>Wikibook</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. Technology Tools, Features, and Resources</td>
<td>Tools for tracking the history of document changes, talk pages, edit pages, hyperlinking, lounges, site statistics, and so forth.</td>
<td>Same tools as Wikipedia but could also use book planning, outlining, and overview tools, enhanced discussion tools, mark-up and commenting tools, enhanced tracking of book contributions and contributors, and ways of fostering collaboration and interaction among those in a Wikibook project.</td>
</tr>
<tr>
<td>10. Qualification and Participant Filters</td>
<td>None, open to all to contribute or edit a document.</td>
<td>None, open to all to contribute or edit a document.</td>
</tr>
<tr>
<td>11. Novices</td>
<td>Wikipedia is more of a collection of articles with random people adding information here and there, then as a way to communicate, collaborate, and protect good work (Bryant, Forte, &amp; Bruckman, 2005).</td>
<td>Wikibooks is a collection of books to which one can contribute to or read.</td>
</tr>
<tr>
<td>12. Experts</td>
<td>Wikipedia is a community of coauthors who use their talents to help establish and maintain the quality of Wikipedia; Wikipedia is more important than any single article or set of articles or resources (Bryant, Forte, &amp; Bruckman, 2005).</td>
<td>Wikibooks is a place wherein one can coordinate a book project that can contribute to the movement of making open knowledge more accessible.</td>
</tr>
<tr>
<td>13. Generation, Distribution, and Maintenance of Ideas</td>
<td>Many to many.</td>
<td>One to many as well as many to many.</td>
</tr>
<tr>
<td>14. Speed of Content Development</td>
<td>Tends to be fast. Assisted by a huge army of volunteer contributors as well as by news and media features or stories. Tends to be fast.</td>
<td>Tends to be relatively slow. Book projects take longer than Wikibookians often expect; many incomplete or unfinished books and resources (Sajjapanroj et al., 2007).</td>
</tr>
</tbody>
</table>
Twin Wiki Wonders?

perspectives, wikis offer common knowledge resources, whether they be dictionaries, encyclopedias, books, or sets of morals, fables, and quotes. At a basic knowledge level, then, wikis provide community-driven access to information resources when needed. From a contributing writers’ viewpoint, on the other hand, wikis are a penultimate example of collaboratively written and socially negotiated texts. Contributions in a wiki can be small pieces of information or media for a presentation, train schedules for a trip to the UK, places to eat at the riverwalk in San Antonio during a conference, and synonyms for a word being defined in Wiktionary. At the same time, they might also be entire chapters or modules of a book or an entire book as in the case of Wikibooks.

As the scope of wiki-work increases from an informational focus at the word, sentence, or paragraph level, as is often the case in Wikipedia or Wikiquotes, to something much more substantive, as in the case of Wikibooks or other online book sites, the problems, issues, and challenges also significantly change. The same wiki tools, approaches, and procedures that work so effectively for Wikipedians may not to be sufficient for Wikibookians. In addition, the level of expertise required to help write a book is at a higher level of specificity and depth than most arriving at a wikibook type of site would likely have. At the same time, the communities of practice that form around such a book have to negotiate issues of chapter coherence, book organization, and topical relevance that those in other types of wiki-related sites do not typically need to consider. Such issues place serious challenges as well as opportunities in the face of wiki researchers as well as designers of new generations of wiki-related technologies.

CONCLUSION

As we pointed out in this chapter, wikis offer a unique window into knowledge negotiation and collaboration processes. The coming decade will undoubtedly witness new tools and resources to track and better understand such negotiation and interaction processes. Now is the time to begin organizing a set of research initiatives, agendas, and questions that will allow for responsive research that helps in the formation of new tools, as well as enhances understanding of the power of wikis. Wikipedia and Wikibooks and many other wiki sites are wondrous tools and resources for twenty-first century working, learning, and living. With these wiki wonders, YOU can now all participate in the human race in new and exciting ways! Engage.

REFERENCES


Twin Wiki Wonders?


ADDITIONAL READING


*This work was previously published in Social Information Technology: Connecting Society and Cultural Issues, edited by T. Kidd and I.L. Chen, pp. 253-272, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 3.18
Wikis as Tools for Collaboration

Jane Klobas
Bocconi University, Italy, & University of Western Australia, Australia

INTRODUCTION

Tim Berners-Lee, the inventor of the World Wide Web, envisioned it as a place where “people can communicate … by sharing their knowledge in a pool … putting their ideas in, as well as taking them out” (Berners-Lee, 1999). For much of its first decade, the Web was, however, primarily a place where the majority of people took ideas out rather than putting them in. This has changed. Many “social software” services now exist on the Web to facilitate social interaction, collaboration and information exchange. This article introduces wikis, jointly edited Web sites and Intranet resources that are accessed through web browsers. After a brief overview of wiki history, we explain wiki technology and philosophy, provide an overview of how wikis are being used for collaboration, and consider some of the issues associated with management of wikis before considering the future of wikis.

In 1995, an American computer programmer, Ward Cunningham, developed some software to help colleagues quickly and easily share computer programming patterns across the Web. He called the software WikiWikiWeb, after the “Wiki Wiki” shuttle bus service at Honolulu International Airport (Cunningham, 2003). As interest in wikis increased, other programmers developed wiki software, most of it (like WikiWikiWeb) open source. Although wiki software was relatively simple by industry standards, some technical knowledge was required to install, maintain and extend the “wiki engines.” Contributors needed to learn and use a markup language to edit pages, and even if the markup languages were often simpler than HTML, non-technical users did not find these early wikis compelling.

In the early years of the twenty-first century, a number of developments led to more widespread use of wikis. Wiki technology became simpler to install and use, open source software was improved, and commercial enterprise-grade wiki software was released. The not insignificant issues associated with attracting and managing a community of people who use a wiki to share their knowledge were discussed in forums such as MeatballWiki (http://www.usemod.com/cgi-bin/mb.pl?action=browse&id=MeatballWiki&olid=FrontPage). The public’s attention was drawn to wikis following the launch, in January 2001, of the publicly written Web-based encyclopedia,
Wikipedia (www.wikipedia.org). And wiki hosting services and application service providers (ASPs) were established to enable individuals and organizations to develop wikis without the need to install and maintain wiki software themselves.

By July 2006, nearly 3,000 wikis were indexed at the wiki indexing site www.wikiindex.org, popular wiki hosting services such as Wikia (www.wikia.org) and seedwiki (www.seedwiki.org) hosted thousands of wikis between them, and Wikipedia had more than four and a half million pages in over 100 languages. Moreover, wikis were increasingly being used in less public ways, to support and enable collaboration in institutions ranging from businesses to the public service and not-for-profit organizations.

**THE NATURE OF WIKIS**

Wiki software allows users to collaboratively edit pages for the Web or intranets. The pages created with wiki software are called “wiki pages” and sites that contain wiki pages are called wiki sites, or simply “wikis.”

Technically, wikis consist of four basic elements:

- Content
- A template which defines the layout of the wiki pages
- Wiki engine, the software that handles all the business logic of the wiki

**Figure 1. How wikis work (Adapted from Klobas & Marlia, 2006)**
Wikis as Tools for Collaboration

- Wiki page, the page that is created by the wiki engine as it displays the content in a browser.

Figure 1 shows how these elements work together.

Wikis consist of pages accessible from a Web browser. They are edited by opening an editing screen in the browser and using either a simple markup language or, increasingly, a rich text editor to edit the text. Links to pages internal or external to the wiki site can be added using simple conventions. These conventions allow a link to be created to a page that does not yet exist; the wiki engine flags such links for future editing. Unless the wiki managers decide otherwise, the content is updated in real time, and once an author saves and closes their changes on the editing screen, the changes are immediately visible online.

Almost all wikis keep track of changes. Older versions of a page can be viewed and, if necessary, restored. Most wikis include a page where recent changes are listed. This feature helps members of the wiki community to keep up to date with changes in content, and can help newcomers get a quick feel for the current concerns of the wiki community. Increasingly, wiki software is integrated with news aggregators like RSS or e-mail notification to alert users to changes without their having to enter the wiki itself.

Another important feature of wikis is the simple permissions structure. Typically, there are three levels of permission: reader, editor, and administrator. Reading permissions may be open to anyone on the World Wide Web or limited to specific, registered individuals. When reading permissions are limited, the wiki is known as a "private wiki."

Various wikis offer support for non-Latin character sets, different media and file types, mathematical notation, style sheets, conflict handling, spam handling, and facilities for merging, exporting and backing up. The different features available in different wiki engines can be seen at the wiki engine comparison site, WikiMatrix (www.wikimatrix.org).

Wiki features are based on design principles established by Ward Cunningham. These principles address human as well as technical goals, for example (in the terms used by Wagner, 2004), wikis are:

- **Open:** If any page is found to be incomplete or poorly organized, any reader can edit it as he/she sees fit.
- **Organic:** The structure and content of the site evolve over time.
- **Universal:** Any writer is automatically an editor and organizer.
- **Observable:** Activity within the site can be watched and revised by any visitor to the site.
- **Tolerant:** Interpretable (even if undesirable) behavior is preferred to error messages.

Wikis usually adopt “soft security,” social conventions that assume that most people behave in good faith, establish that users (rather than the software or a system administrator) operate as peer reviewers of content and behavior, allow that people might make mistakes but that mistakes can be corrected, and emphasize the importance of transparency in their management (Meatball, 2006). Together, these technical features and social principles provide a supportive environment for human collaboration.

**HOW WIKIS ARE BEING USED FOR COLLABORATION**

Wikis are used by groups of people who collaborate to produce information resources that range from meeting agendas to Web sites. They are used in business, government, research, and education.

Public wikis are often built by communities of practice, hobbyists and other interest groups.
Most of these wikis are concerned with quite specific topics such as a specific sport, book, author, religion, or philosophy. Often, they are maintained and read by small groups of friends and colleagues. An example is ukcider, a wiki that provides information about real cider (http://ukcider.co.uk/wiki) and supports advocacy for real cider enthusiasts and small producers.

Wikis can be used to create and maintain knowledge repositories for communities of practice (Godwin-Jones, 2003; Roberts, 2005). Members of the community update the repository with information and solutions to problems. As with other types of wiki, the communities served might be public or private. Potential uses of private wikis to increase autonomy among US intelligence workers were described by Andrus (2005).

Committees and working groups, particularly those that work across institutional and/or geographical boundaries, use wikis to develop and maintain agendas and documentation. The DCMI Education Working Group Wiki (http://dublincore.org/educationwiki), for example, contains work plans, FAQs, agendas, notes, documentation, and links to key references for the group.

A common organizational use of wikis is project support (Bean & Hott, 2005). Wikis are used to support the sharing of agendas, ideas, resources, plans and schedules. Documents can be jointly produced within the wiki or made available as attachments. Angeles (2004) describes how wikis have been used by Lucent Technologies for project documentation including preparation of meeting notes, product specification notes, product requirements documents, project deliverables, content audits, technical documentation, and style guides. Other companies that use wikis for project support include Michelin China and the investment bank, DrKW (Paquet, 2006).

Wikis are also being used to write and maintain system documentation. A public wiki has been developed by Mozilla, the provider of the Firefox Web browser, to maintain Mozilla product documentation (http://kb.mozillazine.org). Any member of the public can contribute. Some companies have established private wikis to maintain documentation for their internal information systems (IS). Users contribute to documentation based on their experience with the system.

Organizations and groups can use wikis to produce reports of news or events. Schools, in particular, appear to be experimenting with this type of wiki, in which parents, teachers, administrators, and students can all contribute reports, photographs, audio and video of a school event.

Some groups use wikis to plan conferences or meetings and to continue discussions that were begun at a meeting. Examples include the American Library Association (ALA) 2006 Conference wiki (http://meredith.wolfwater.com/ala2006/) and the Yale Access to Knowledge wiki (http://research.yale.edu/isp/a2k/wiki/). Some experiments in use of wikis during conferences have found them valuable (Boyd, 2004; Suter, Alexander, & Kaplan, 2005). Boyd describes how workshop participants quickly learnt to use wikis to capture “their own streams of consciousness or the comments of others.”

Wikis are used in a variety of ways for collaboration in classrooms ranging from primary school to university (Bold, 2006; Ferris & Wilder, 2006; Lamb, 2004; Skiba, 2005). Mitchell (2006) notes that wikis are consistent with modern theories of learning such as “connectivism” (Siemens, 2004), the idea that learning occurs in a complex and changing environment that cannot be controlled. In the classroom, wikis are used in collaborative writing, student projects and group assignments. Collaborative writing tasks can be given to develop writing and collaboration skills or to develop information resources for use by other students.

Wikis are particularly well adapted to shared creation of directories and lists such as bibliographies, staff lists and job vacancies. In schools and universities, teachers who use wikis in this
way assign groups of students or whole classes the task of jointly preparing annotated lists of information resources. Although most educational wikis are private, some become public resources. An example is the Wiki for IS Scholarship maintained by the Information Systems Department in the Weatherhead School of Management at Case Western Reserve University (http://isworld.student.cwru.edu/tiki/tiki-index.php) which contains summaries of major books and articles in IS as well as reviews of the contributions of major IS scholars. Often, wiki-based directories also provide access to resources. Participants in the Government Open Code Collaborative (http://www.gocc.gov) use a public wiki to list and describe source code that can be shared among its members.

Wiki software is also being used to create Web sites that draw on the contributions of a number of people. One example is the site developed by the Bach-Academie of Montreal to publicize a series of concerts. The musicians themselves added details to the wiki, which was then published to the Web (http://www.bach-academie-de-montreal.com/).

ISSUES IN WIKI MANAGEMENT

The ease with which individuals can contribute to wikis, the open nature of contribution, and the philosophy that the guardians of the content are the people who create the wiki can produce outstanding knowledge resources. On the other hand, these very characteristics can also be associated with problems such as reluctance to contribute or chaos through unstructured contribution. Some management is therefore required by the person or group that wishes to develop and maintain the wiki. The most significant management issues tend to be those of encouraging people to contribute, and managing intellectual property, problem users and spam.

As with any knowledge resource, people need both a reason to contribute to a wiki and a sense that they have the ability to do so. Wikis created for small groups or communities with a specific purpose should quite readily attract users if those users also know how to access and use the wiki. For example, a wiki established for the purpose of jointly preparing an agenda or document, where all members of the group responsible for preparing the document are motivated to contribute and know how to access the wiki and use the editor, is likely to be used by the intended users for the intended purpose. The more diffuse the purpose and the user group, the more uncertainty there will be about the development and continuation of a wiki. In practice, however we usually find a core group of regular contributors and a peripheral group of occasional contributors and people who read the wiki but do not contribute to it. (This pattern is common in online communities; see, for example, Wenger, McDermott, & Snyder, 2002). The core users can act as role models for other users. Other patterns that have been observed include the champion pattern (“single wiki-nut, encourages coworkers to add, view, improve”) and the trellis pattern (“egregiously boring content calls for fixing ... My [contribution can be] far more interesting than that!” Confluence, 2006).

Techniques for ensuring that potential contributors feel able to contribute include providing simple, easy to follow documentation for editing pages and assurance that mistakes are permitted and can quickly be remedied. Anti-patterns that discourage use include the gate pattern, “too many procedural barriers to adding content” (Confluence, 2006).

Where the content of a wiki may include information, images and files drawn from other sources, the intellectual property of the original sources must be respected. Standard policies for dealing with material that is the intellectual property of others include: (1) only material that does not
violate the copyright of the original creator can be included in the wiki and (2) the sources of all included material should be acknowledged. Equally, a decision has to be made about ownership and conditions for re-use of original material included in a wiki. The fact that the content of any single wiki page is usually produced by multiple authors needs to be acknowledged. A typical statement of the intellectual property in a wiki acknowledges its joint authorship and permits re-use, provided the wiki page is acknowledged as the source of the material. The GNU Free Documentation License used by Wikipedia provides an example (Free Software Foundation, 2002).

The more open a wiki to the public, and the larger the wiki user community, the greater the potential of encountering a problem user. Problem users may post unacknowledged copyrighted material, post material designed to offend or anger readers or contributors, remove material or otherwise “vandalize” the wiki. Some wikis use registration to limit the possibility that a problem user will continue to damage the wiki, but many wiki communities prefer to adopt the soft security principles described earlier in this article. It may be sufficient simply to let someone know (privately, by e-mail) the first time they post unacceptable content, that this type of content is not appropriate for the wiki (Turnbull, 2004). A wiki page that contains problem content can easily be replaced by an earlier, “clean” version. The active members of a wiki community often share the task of watching pages and dealing with problems.

Another potential problem, for public wikis in particular, is spam. Pages can be watched for spam in the same way that they are watched for other problem content, but this can be a time-consuming task. Some wikis use registration processes that require human intervention, such as CAPTCHA (Carnegie Mellon University School of Computer Science, 2000-2005), to prevent mass spam attacks. Spam capturing services are being added to wiki engines to reduce the amount of spam that reaches a wiki page.

FUTURE TRENDS AND CONCLUSION

At the time of writing, wikis were still the domain of early adopters, rather than a part of the mainstream. Despite widespread use of Wikpedia, knowledge of what wikis are and how they can be used for collaboration, particularly for private use, is not widespread. As of the end of 2006, only two non-technical books about wikis had been published (Klobas, 2006; Tapscott & Williams, 2006). Nonetheless, in October 2006, Google Inc. bought the wiki software and Web hosting service, JotSpot, an indication that consumer use of wikis is on the rise. The Gartner Group envisages wikis reaching maturity for business use some time before 2010 (Fenn & Linden, 2005). As wiki software matures, we can expect improvements in technical qualities such as ease of implementation and stability, as well as improvements in editing interfaces and graphical quality. These improvements, along with the positive experiences of early adopters, should help gain the interest and confidence of potential users and result in further diffusion of wikis as tools for collaboration.

REFERENCES


Skiba, D. J. (2005). Do your students wiki? *Nursing Education Perspectives, 26*(2), 120-121.


KEY TERMS

Editor: The authors of wiki pages may be called “editors” because they have editing permissions.

Soft Security: Social conventions for trust, peer review and correction of errors adopted by contributors to wikis.

Wiki: An information resource that is created by multiple authors who use Web browsers that interacts with wiki software.

Wiki Engine: The software that handles the business logic of a wiki.

Wiki Page: A page of wiki content displayed in a Web browser.

Wiki Site: A set of related wiki pages. When a wiki site can be viewed on the World Wide Web, it is also a Web site.

Wiki Software: The suite of software used to produce and manage a wiki. This software may include, in addition to the wiki engine, add-ons and extensions that extend the feature set and functionality of the wiki.

WikiWikiWeb: The first wiki engine, written by Ward Cunningham.

This work was previously published in the Encyclopedia of E-Collaboration, edited by N. Kock, pp. 712-717, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
INTRODUCTION

Although scientific research has always been a social activity, in recent years the adoption of Internet-based communication tools by researchers (e.g., e-mail, electronic discussion boards, electronic mailing lists, videoconferencing, weblogs) has led to profound changes in social interaction and collaboration among them. Research suggests that Internet technologies can improve and increase communication among noncollocated researchers, increase the size of work groups, increase equality of access to information by helping to integrate disadvantaged and less established researchers, help to coordinate work more efficiently, help to exchange documents and information quickly (Carley & Wendt, 1991; Nentwich, 2003). There is abundant research on new forms of group work originated from the use of computer technologies. Carley and Wendt (1991) use the term extended research group to refer to very large, cohesive, and highly cooperative research groups that, even being geographically dispersed, are coordinated under the supervision of a single director. The term collaboratory is also used to refer to similar groups (Finholt, 2002). Although there is much research on how Internet technologies are used by unified and cohesive work groups to collaborate (e.g., Moon & Sproull, 2002; Walsh & Maloney, 2002), less attention has been paid to how the Internet facilitates collaboration among researchers outside these highly cohesive groups. Weblogs (blogs) can become a useful tool for this type of collaboration and for the creation of virtual groups. Weblogs are frequently updated Web pages, consisting of many relatively short postings, organized in reverse chronological order, which tend to include the date, and a comment button so that readers can answer (Herring, Scheidt, Bonus, & Wright, 2004). They enable users to communicate with a worldwide nonrestricted community of people in similar fields, which leads to several forms of collaboration. The purpose of this article is to
present a brief overview of the uses of weblogs as tools for research e-collaboration.

Defining the concept of “research e-collaboration” precisely is extremely difficult. Here we assume that members of a virtual community engage in research e-collaboration when they use e-collaborating technologies in order to share information and discuss issues which contribute to advancing knowledge in a specific area.

BACKGROUND

The term Weblog was coined by Jorn Barger in 1997 to refer to personal Web sites that offer frequently updated information, with commentary and links. Blood (2002) classifies blogs into two “styles”: the filter type, which includes links pointing to other sites and comment on the information on those sites, and the personal-journal type, with more emphasis on personal self-expressive writing. There are many other types of blogs described in the literature, defined on the basis of different criteria; for example, knowledge blogs (k-blog), community blogs, meta-blogs.

The capabilities of blogs make them helpful tools for communication between members of a community or organisation. Some types of weblogs have originated as an answer to the communicative needs of specific communities; for example, knowledge blogs, weblogs for personal knowledge publishing. Kelleher and Miller (2006) describe “knowledge blogs” as “the online equivalent of professional journals” used by authors to document new knowledge in their disciplines. A related concept is that of “personal knowledge publishing,” defined by Paquet (2002) as “an activity where a knowledge worker or researcher makes his observations, ideas, insights, interrogations, and reactions to others’ writing publicly in the form of a weblog.” Many corporate and academic blogs make use of capabilities that afford collaboration: they enable scholars to communicate with a wide community, fostering peer review and public discussion with researchers from different disciplines. These weblogs have a precedent in what Harnard (1990) terms “scholarly skywriting”: using multiple e-mail and topic threaded Web archives (e.g., electronic discussion) to post information that anybody can see and add their own comments to.

There are many types of academic blogs (blogs from journal editors, individual scholars’ blogs, research groups’ blogs, PhD blogs), each of them used for different purposes. For instance, while the main purpose of the weblogs implemented by universities is discussion, weblogs by PhD students are mainly used to comment on the day’s progress and on the process of PhD writing, and blogs from journal editors are usually filter blogs, which provide links to articles or which comment on news related to the journal topic.

The uses of weblogs in research have been discussed in several papers and blog posts (Aïmeur, Brassard, & Paquet, 2003; Efimova, 2004; Efimova & de Moor, 2005; Farmer, 2003; Mortensen & Walker, 2002; Paquet, 2002). These researchers depict blogs as facilitating scientific enquiry in two ways: (1) they help to access and manage content, through features such as archives, RSS (an automated system that enables bloggers to syndicate their content to other blogs), searchable databases, post categories; and (2) they are tools for collaboration, through communication and network features. These features include hyperlinks, comments, trackbacks (records of the Web address of the blogs that have linked to a blog posting), RSS, or blogrolls (a list of blogs the author usually reads, and that, therefore, deal with similar topics). But the most important “ingredient” for collaboration is the bloggers’ perception of blogs as a means to point to, comment on, and circulate information and material from other blogs.

As a tool for collaborative research, blogs have several uses: (1) supporting community forming; (2) helping to find other competent people with relevant work; (3) facilitating connections
between researchers: blogs make it easier for researchers to develop, maintain and activate connections with others; (4) making it possible to obtain speedy feedback on ideas; (5) fostering diversity and allowing for radical new ideas to be acknowledged, circulated and discussed; (6) fostering communication and collaboration between researchers from different fields; (7) supporting the initiation and development of conversations. In weblogs, ideas can be generated and developed through discussion with others.

**E-COLLABORATING THROUGH ACADEMIC WEBLOGS**

We used a corpus of 100 academic English-language weblogs, collected between January and March, 2006, in order to analyse how they are used as tools for e-collaboration and how the elements of weblogs contribute to this purpose. Many academic weblogs are not intended as collaborative tools and thus do not include social software. In other cases, although comments are allowed, interactivity is very limited, and the weblogs are not really used as conversation tools. Therefore, for the purpose of this research, we have selected only active weblogs that include and make use of the comment feature. Drawing on previous research on blog analysis (e.g., Herring et al., 2004) and on our initial inspection of the blogs of the corpus, we identified the functional properties that make interaction and collaboration possible (e.g., comments, links). Then, we analysed the function of the 15 most recent entries in each blog (e.g., ask for feedback), the function of the comments to these entries (e.g., disagree with a post), and the types of links included in the postings and comments (e.g., to other blogs, to news sites, to Web sites by others, to the blogger’s own site).

The analysis showed that academic weblogs can be used to support the following forms of collaboration (or collaboration stages):

1. **Finding like-minded peers and establishing contact and collaboration with them, thus helping to create a virtual community of people interested in the same topic.**

Weblogs share features of personal and public genres. As personal genres, they are used by researchers to keep records of their thoughts, ideas or impressions while carrying out research activities. They also have an open nature, since anybody can access and comment on the author’s notes, and they are in fact intended to be read. By revealing the research a researcher is currently involved in and, at the same time, providing tools for immediate feedback on that research, weblogs facilitate collaboration. In addition, since weblogs are open to the public, not restricted to peers working in the same field, they facilitate contact among scholars from different disciplines and invite interdisciplinary knowledge construction (Aïmeur et al., 2003).

Blogrolling lists enable readers of a weblog to find peers with relevant work. These lists do not only point to related work, but also function as signs of personal recommendation, which help others to find relevant contacts faster (Efimova, 2004). Links and comments are also useful to expand a researcher’s community. Other researchers read the postings in the research blog and leave comments, usually including links to their own research. Authors of weblogs can also include links to articles on similar work carried out by other researchers, thus enabling the readers to get into contact with them.

Many researchers use the comment feature to explain that they are working on areas similar to the ones dealt with in the original posting and to suggest the possibility of collaboration or contact, in a more or less direct way (e.g., “I would like to keep in touch about your findings”). Community blogs are sometimes used by bloggers to summarize briefly their research and invite collaboration and contact with others.
Academic Weblogs as Tools for E-Collaboration Among Researchers

a. I am in the middle of a research project that looks at (...). I have a research log in which I make occasional observations, related and sometimes unrelated, to the ongoing work. Interested? Let me know and I will give you the URL.

The entry above got several comments of people eager to exchange ideas; for example:

b. I'd love to exchange ideas (...). How can we establish contact?

2. Asking for and receiving advice/feedback from others.

Weblogs enable researchers to post an idea or theory and have it circulated and discussed, thus helping the author revise and refine it. Authors who want to use weblogs for this purpose embed elements that facilitate hyperlinked conversations into them, such as links, comment software, and trackback features.

Very frequently, authors use weblogs to explicitly request feedback, comments, suggestions, and so forth on a paper, presentation, or idea. Comment software facilitates the discussion between the researchers asking for feedback and anybody wanting to contribute: Other bloggers state their agreement and disagreement with the author and with other commentators, ask questions, offer new perspectives, and so forth.

a. I was asked by the editors of the New Palgrave Dictionary of Economics to contribute a short article on the analysis of variance (...). Here's the article. Any comments (...) would be appreciated.

The entry above got eight comments suggesting ways to improve the paper; for example:

b. * Missing from the draft is a discussion of ANOVA versus regression (...)

* Here is one question you might address: When is ANOVA preferred to linear regression? (...)

Comments may also include links to other blogs with similar topics and thus help the original author with other people's perspectives. The weblog becomes a forum for discussion of the paper prior to publication, similar to that of a conference. When a blogger requests and gets feedback on a paper, he/she usually acknowledges the help from the others and shows the result of the revision in an update to the original entry. The result is a more consensual paper, which makes use of knowledge from other researchers.

Posting articles for peers to comment on in a weblog has several advantages. In the first place, the author can get immediate and public feedback, thus accelerating the development of new knowledge and involving many people in the review process: Anybody can contribute, even people unknown by the author, nonestablished researchers who are not usually involved in peer-reviewed publishing, and researchers from different cultural backgrounds. Besides, since bloggers usually comment on other's ideas on their own blogs, feedback may come from others than just immediate readers. Authors can post any idea, even very innovative or radical ideas which would not be accepted in traditional publishing, and, as Paquet (2002) points out, they can use weblogs to let others know not only about positive results, but also about negative results, blind alleys and problems. Peers can help by providing their ideas, suggestions, and so forth, and this way knowledge construction becomes a communal activity.

In many cases, researchers do not ask for feedback, but for references, information or resources on a topic, or help to solve a problem, hence using blogs as a tool to get information on a topic from competent peers they may not know personally.

3. Discussing topics or theories in which several researchers are interested, thus
Academic Weblogs as Tools for E-Collaboration Among Researchers

contributing to developing or articulating an idea and to reinforcing the links between the members of the virtual community.

Efimova and de Moor (2005) show how weblogs can serve as conversation tools which support the exchange of multiple perspectives and fast and meaningful reactions and facilitate the joint development of ideas. They define a new form of conversation enabled through the use of weblogs, “distributed weblog conversation”: conversations across many weblogs, where several authors link their ideas to construct a collective piece of knowledge.

The joint articulation of ideas by several researchers relies on elements such as links, comments and trackback features. Links may be to text but also to audio or video documents presenting or illustrating an idea, to simulations, to other Web pages or to other posts. Scholars use links to present and discuss an idea by someone else, or to articulate and summarize a previous discussion, before contributing their own idea and presenting it for others to comment on. Very frequently, when presenting others’ arguments on a topic, scholars include quotes accompanied by links to the original post, that way allowing the reader to follow the conversation in the original blog.

The comment feature facilitates discussion based on the posting. Bloggers write their comments, get answers back from the author of the original post, comment on other bloggers’ comments, and so forth, thus engaging in a conversation that may contribute to the development of innovative ideas. Comments are used for many different purposes: to show agreement or disagreement with the blogger, to ask for clarification on a specific point that the commentator does not understand or does not agree with, to request further information, to strengthen a point, to provide links to one’s own blog, and so forth. This back-and-forth may lead to collaborative revision and refinement, and even permuting, of an idea or argument, e.g.:

a. * In general, I really side with your post. Over at Anthropology.net, there have been a couple posts and discussions on (...). However, this post of yours, much more eloquently advocates breaking the boundaries between the West and non-West (...)

* Thanks, X. I’ll have a look at the posts you link to (…)

Since discussion of an idea may be distributed across several weblogs, it is necessary to have mechanisms to keep track of the discussion that one’s posting may have sparked. One such mechanism is trackbacks. Trackback is a special feature of the program Movable Type, which notifies the blogger when another weblog has linked to his/her weblog. Other users can view the trackbacks and thus follow the discussion in other weblogs.

FUTURE TRENDS

Weblogs are a genre whose capabilities offer a great potential for collaboration among researchers. However, there are still few academics who engage in blogging, and in many cases academic blogs are used for purposes other than collaboration. Several reasons have been suggested: the risk of sharing information and having ideas stolen or attacked before the research is published, the fear of damaging credibility, the time that blogging takes away from more traditional research activities (Lawley, 2002; Mortensen & Walker, 2002).

Since the weblog is quite a recent genre, there is little research on how weblogs are used for enquiry and knowledge creation. Their open nature makes weblogs appropriate for nonestablished researchers, who can that way collaborate with peers they don’t know, and researchers who seek collaboration with a worldwide non-discipline-restricted community of researchers. There are many different types of weblogs written by
academics and researchers. Further research is needed to determine which of these weblogs are more involved in collaboration, which form of collaboration takes place in the different academic weblogs, and how bloggers use language to engage in collaboration.

CONCLUSION

In this article, we start from the assumption that, since scientific knowledge is socially constructed, scientific knowledge creation is collaborative in nature. Thus, rather than considering collaboration as restricted to the work by unified and cohesive groups involved in a project under the supervision of a coordinator, we see it as any sharing of information and exchange of perspectives which contribute to advancing knowledge in a specific area.

Weblogs are a genre whose capabilities allow the exchange of any type of information and support conversations with a rhizomatic structure, thus being an appropriate tool for collaboration and for the sharing of ideas among several researchers in different places around the world. Weblogs can represent and distribute knowledge in different forms; that is, text, multimedia, simulations. They have a modularized format, relying highly on hypertext: weblogs consist of different postings, linked together to postings in other weblogs, on which they comment or which comment on them, making it possible to construct a web of knowledge on a topic. Links and comments allow connecting different pieces of related research into a coherent whole, thus facilitating the development of cumulative knowledge.

The use of CMC tools (and specifically of weblogs) does not only lead to an increase in collaboration, but also to a change in collaboration patterns. Weblogs enable almost instantaneous communication and feedback among a worldwide community, thus allowing for a much more interactive generation and articulation of ideas and theories and for the constant revision and update of theories. The informal and open nature of weblogs facilitates collaboration among anybody interested in a topic and, hence, fosters interdisciplinary collaboration.

REFERENCES


Farmer, J. (2003). Personal and collaborative publishing (PCP): Facilitating the advancement of online communication and expression within
Academic Weblogs as Tools for E-Collaboration Among Researchers


KEY TERMS

Blogroll: A list of links to Web pages the author of a blog finds interesting.

Comment: A link to a window where readers can leave their comments or read others’ comments and responses from the author.

Personal Knowledge Publishing: The use of Weblogs by knowledge workers or researchers to make their observations, ideas, insights and reactions to others’ writing public.

Research E-Collaboration: The use of e-collaborating technologies in order to share information and discuss issues which contribute to advancing knowledge in a specific area.

Scholarly Skywriting: Using multiple e-mail and topic threaded Web archives (e.g., electronic discussion) to post information that anybody can see and add their own comments to.

Trackback: A link to notify the blogger that his/her post has been referred to in another blog.

Weblog: A frequently updated Web page, consisting of many relatively short postings, organized in reverse chronological order, which tend to include the date and a comment button so that readers can answer.

This work was previously published in the Encyclopedia of E-Collaboration, edited by N. Kock, pp. 1-6, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
INTRODUCTION

Education is one of the key sectors that has benefited from the continuous developments and innovations in information and communication technology (ICT). Web-based facilities now provide a medium for learning and a vehicle for information dissemination and knowledge creation (Khine, 2003). Accordingly, developments in ICTs provide opportunities for educators to expand and refine frameworks for delivering courses in innovative and interactive ways that assist students achieve learning outcomes (Kamel & Wahba, 2003). However, the adoption of ICTs has also created tensions between traditional control and directiveness in teaching and student-centred learning, which relies on flexibility, connectivity, and interactivity of technology-rich environments.

This chapter examines the introduction of Web-based technologies within a media studies course. The objective was to establish a community of learning, which provides students with a portal or entranceway into a common work area and out to networks of media-related organizations. So doing, a pilot study was conducted within the Department of Communication at Texas A&M University to blend Weblog facilities with a classroom setting to enhance students’ interpersonal and content interaction, and build citizenship through participation and collaborative processes. Four key aims frame this study:

1. provide an accessible, interactive online environment in which students can participate with peers and engage with new media technologies within a learning community setting;
2. develop an instructional technology framework that enhances the learning experience and outcomes within online educative environments;
3. establish a portal or gateway for students to access media advocacy and special interest groups and enhance and diversify perspectives on global media; and
4. evaluate student-learning experiences facilitated through innovative online instructional technologies.

BACKGROUND

Early approaches to integrating ICTs into education environments emerged from conventional learning models, originating from the objectivist approach in which a reality exists and experts instruct individuals of that reality (Belanger & Slyke, 2000). However, such teacher-centric, information-based approaches failed to adequately prepare students to become independent learners. Responding to these limitations, educators embraced learner-centric approaches such as constructivism, which leaned weight to the empowerment of individuals to take charge of their own learning environments. As Wilson (1996) suggests, the constructivist movement in instructional design emphasized the importance of providing meaningful, authentic activities that can help the learner to construct understandings and develop skills relevant to solving problems and not overloading them with too much information. Solis (1997) supports this position, suggesting that student-centred learning “… relies on groups of students being engaged in active exploration, construction, and learning through problem solving, rather than in passive consumption of textbook materials” (p. 393).

In spite of these favorable positions, Khine (2003) warns that creating such learning environments supported by ICTs can be intrinsically problematic. Accordingly, it is critically important that careful planning and design is employed at the early stages of instructional design to provide proper support and guidance, as well as rich resources and tools compatible to each context. When adequate consideration is given to new learning and teaching strategies that incorporate ICTs, real opportunities exist for educators to provide students with a dynamic environment to learn, to think critically, and to undertake productive discussions with their peers in supportive, constructive environments. Given the potential of such technology-rich learning environments, educators have the opportunity to make student learning more interesting and enriching, preparing them for the demands of the future workplace. Accordingly, instructional designers must consider matching the strengths of new technology (flexibility, connectivity, and interactivity) with traditional forms of education (control and directiveness) to inspire, motivate, and excite students in ways that maximize the individual’s learning potential.

Achieving these goals requires the development of individual competencies in problem solving, participation, and collaboration, and communities of learning (Kernery, 2000; Khine, 2003; Wilson & Lowry, 2000). Problem solving provides ways for students to engage with authentic episodes, providing opportunities for students and educators to examine events and reflect on solutions. One way of maximizing the benefits of problem solving is to support these through collaborative processes, which can be built around these “episodes” by focusing on the use of instructional methods to encourage students to work together as active participants on such tasks. Such efforts can be facilitated through structuring and organizing online interactions using computer-mediated communication, which provides the means to overcome limitations of time and place (Harasim, Calvert, & Groeneboer, 1997). Based on the principles of the transformative paradigm, multiple perspectives, and flexible methods, it is possible for students to adapt, to process and to filter content into their own logical frameworks, resulting in outcomes that may not be thoroughly predictable (Bento & Schuster, 2003). As Morrison and Guenther (2000) note, such collaborative environments provide a forum for students to discuss issues, engage in dialogue, and share results. However, Bento et al. (2003) also warn that one of the main challenges in Web-based education...
is to achieve adequate participation. They offer a four-quadrant taxonomy of learner behaviours when participating in interpersonal and content interaction--missing-in-action, witness learners, social participants, and active learners—as a way of understanding this dynamic.

When building components of problem solving, participation, and collaboration around small group processes, or learning communities, it is critical that these dynamics have beneficial effects on student achievements and psychological well-being. As Khine (2003) suggests, building community imparts a common sense of purpose that assists members grow through meaningful relationships. Accordingly, learning communities can be “… characterized by associated groups of learners, sharing common values and a common understanding of purpose, acting within a context of curricular and co-curricular structures and functions that link traditional disciplines and these structures” (p. 23). Rickard (2000) equates the notion of common values to that of “campus citizen,” in which students not only engage in educative endeavour but also learn networking and become “life-long members of our communities” (p. 13).

Such communities though should not be limited to just participants within narrowly defined student groups or the domain of institutional environments. ICTs also provide opportunities for students to connect to other informative communities related to their area of discipline or study via the Web. So doing, the educator can build into the instructional design gateways or portals to direct participants of small learning communities to other relevant organizations, groups, and individuals to extend campus citizenry and enhance knowledge creation. Tatnall (2005) suggests that such portals can be seen:

... as a special Internet (or intranet) site designed to act as a gateway to give access to other sites. A portal aggregates information from multiple sources and makes that information available to various users .... In other words, a portal offers centralised access to all relevant content and applications. (pp. 3-4)

In accessing Tatnall’s (2005) notion of portals, it is possible to think of these starting points in diverse ways. While no definitive categorization of the types of portals exists, Davison et al. (2003) offer a list of possible alternatives: general portals, community portals, vertical industry portals, horizontal industry portals, enterprise information portals, e-marketplace portals, personal/mobile portals, information portals, and niche portals. However, it is important to point out that these categories are not mutually exclusive, highlighting the malleable nature of these gateways for educators to blend the strengths of one with advantages of others to achieve a more effective portal design.

Even though portals are conceptually difficult to categorize and define, there exist a number of important characteristics that assist in facilitating the objectives of gateways as access points to, and aggregators of, information from multiple sources. For example, Tatnall (2005) draws from a number of scholars to present a general guideline of beneficial characteristics employed to facilitate community collaboration amongst users and the rapid sharing of relevant content and information. For example, the characteristics of access, usability, and functionality, as well as sticky web features like chat rooms, e-mail, and calendar functions, have been used to assist in maintaining user interest, attention, and participation within a site.

Within the education sector, portals offer great potential to achieve the kinds of goals laid out so far. However, portal development in these learning environments present a number of challenges for universities as they continue to grapple with a variety of business, organizational, technical, and policy questions within the framework of a larger technology architecture. For example, Katz (2002) highlights the following challenges:
build standards to create compelling and “sticky” Web environments that establish communities rather than attract surfers;
• create portal sites that remain compelling to different members of the community within the short-term and throughout student lives;
• create the technical and organizational infrastructure to foster “cradle-to-endowment” relationships via virtual environments; and
• integrate physical and virtual sites to foster social and intellectual interactions worthy of the university’s mission.

Within this framework of challenges, the emphasis continues to return to the realization that bringing education and ICTs together strategically relies less on technology per se than on the educator’s ability to design portals that are instructionally sound (to meet learner needs), customizable and personalized (for each member of the community), meets the institutional vision and mission, and fits within the larger technology infrastructure. However, Katz (2002) argues that while the challenges are great, opportunities can be realized. With universities engaged in relationship management enterprises, the situation requires a proactive role in developing a “…belief system, a worldview, a set of approaches and technologies organized to arrange and rearrange both our front door and our rich holdings” (p. 13).

To achieve and maintain such connections, TAMU implemented a three-tier model for development and implementation of foundation, cornerstone, and capstone learning communities within the university setting (see Figure 1). In developing this model for establishing learning communities, a number of recommendations were made:

1. Draw from faculty and staff knowledge to create initiatives that make existing educational opportunities even better.
2. Provide resources and incentives to develop, implement, and access learning communities; and encourage innovative pedagogies to support and reward the implementation of technology-mediated instruction, use of peer instructions and collaborative learning.
3. Provide a clearinghouse so the benefits from experience and expertise can be utilized in future design and implementation of learning communities. (TAMU, 2005)

With this framework in mind, a Weblog was integrated into the COMM 458 Global Media course using a generic blogger.com template as part of pilot study funded by a $3,000 TAMU Summer Institute for Instructional Technology Innovation Grant. The template included an archive (of posts and comments), calendar function, a database of user information including e-mail links to registered members, e-mail notification facility, and links to other relevant websites. In providing 22 students access to Weblog facilities,
the study set out to establish a model of content production and platform interoperability for inspiring innovative and creative appropriation of Web-based facilities by TAMU students and faculty. Accordingly, the Weblog was strategically integrated into the course to achieve the following outcomes:

- facilitate community building and understandings of social networking and citizenship at local, national and international levels;
- provide a platform for students to evaluate, communicate and critique current issues and problems within informal and engaging online environments;
- improve content, information and knowledge management within the course structure;
- improve teaching and learning through understanding the dynamic between participation and collaborative processes within a blended educative environment; and
- develop alternative pedagogical and technological strategies to support and encourage each student to become an independent learner.

All students were invited to be members of the Weblog community by e-mail, which contained a hyperlink that directed them to blogger.com’s registration site. There, students selected a user name and password, which brought them into the course Weblog site (see Figure 2). Once in the site, members could interact with each other by posting summaries of readings (including discussion points and questions), comment on posts in which

<table>
<thead>
<tr>
<th>Time Line</th>
<th>Courses</th>
<th>Content</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>Foundation</td>
<td>Introduce competencies</td>
<td>Critically analyze, Personal integrity, Contribute to society, Communication</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core Curriculum Courses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sophomore, Junior, or both</td>
<td>Cornerstone</td>
<td>Reinforce and integrate competencies in the discipline</td>
<td>Critically analyse, Master depth of knowledge, Communication</td>
</tr>
<tr>
<td>Senior</td>
<td>Capstone</td>
<td>Emphasize, synthesize and apply competencies in the discipline</td>
<td>Critically analyse, Master depth of knowledge, Personal integrity, Communication, Contribute to society</td>
</tr>
<tr>
<td>Career or Graduate School</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Assessing Weblogs as Education Portals

students raised issues and problems (moderated by the course instructor), and source stories and information from national and international news services (e.g., CNN, BBC, or Aljazeera). Users were also given the task of building a database of relevant links to outside media organizations and advocacy groups to assist each other in answering problems and issues posed within online and classroom discussions. All students were assigned administration rights to the Weblog to instil a sense of ownership among members.

FUTURE TRENDS

Participation within the Global Media course (online and classroom) was measured in several ways, including the number of entries (posts and comments), user log-ins, survey, interviews, and observations. Overall, students and instructor made a total of 303 entries (72 posts and 231 comments) over the 10-week period from September to November of the 2005 Fall Semester. On average, there were 3.25 comments per post. However, included in this number were 108 comments (81% response rate) by students to a series of six quizzes (which accounted for 15% of the course’s total grade) implemented throughout the semester to test students’ general knowledge of key theories and concepts. When the quiz comments were excluded and the figure re-calculated, the average number of comments per post significantly decreased to 1.70. These figures would suggest a relatively poor level of participation in online interactions. However, these figures do not reveal the complexity of the online-classroom dynamic in catering for different learner preferences. Another way of assessing student participation
is to categorize their behavior into types of interactions through observing participation in both settings.

Drawing from Bento et al. (2003) four-quadrant taxonomy, it is possible to map the types of behaviors these learners engaged in as part of the participatory and interactivity components of the Global Media course (see Figure 3). Quadrants I and II share the characteristics of low interpersonal interaction. However, whereas Quadrant I students do not care about course content and their peers, “witness learners” are actively engaged with the course materials and discussion (high content interaction), log in frequently and do all the readings but do not actively contribute to course discourse (low interpersonal interaction). On the other hand, Quadrants III and IV share the characteristic of high interpersonal interaction. Students in Quadrant III (high interpersonal interaction, low content interaction) thrive on...
the social aspects of education, demonstrating characteristics of great conversationalists, with high communication and interpersonal skills. Such students, though, engage in social interaction most often at the expense of reflection and thoughtful consideration of course content. Quadrant IV represents what educators define as “good participation,” with high content interaction and interpersonal interaction. Their contributions in discussions are substantive and frequent. Accordingly, Bento et al. (2003) argue that these students contribute not only to the problem-solving task but also to building and sustaining relationships in the learning community.

Given the findings of the study, the Weblog added value to the learning community by providing students with access into a common area to connect and interact with one another and as a gateway out to sources of information beyond the confines of the institutional environment. Through student-established links to media organizations and advocacy groups, members connected to important and divergent understandings and knowledge of the role of media in a global context. One student comment encapsulates the general feeling of course participants toward the Weblog facilities:

... Even if all students had equal enthusiasm to participate in class discussion, class time limits everyone from making significant contributions and feeling part of learning. Blogs provide a place for sharing related material [discussion and provision of links to outside sources] that the classroom setting does not, and a way to question or comment on material we have already covered. In general, blogging let’s students engage with each other in a way different from class. By allowing discussions to continue and expand, it helps students who struggle with the material or who are afraid to ask questions in class. It certainly helped me to learn much more than just reading textbook information.

In spite of these positive responses, the chronologically organized Weblog structure offered insufficient flexibility and functionality to increase online participation. Given the findings, the study recommends re-designing the Web-based learning environment using software, innovative interfaces, and workflows. So doing, there is a need to:

1. Develop a technology interface that acts as an entranceway into a series of linked, embedded pages so (a) information storage and retrieval; (b) discussion and commentary; (c) communication; and (d) project facilitation can be purposefully compartmentalize to improve technology, content, interpersonal and intellectual interaction.

2. Integrate chat room facilities into the portal design to create more compelling, sticky Web-environments for participants to interact in real-time with peers and instructor.

3. Introduce more comprehensive monitoring facilities to track technology, content, interpersonal and intellectual interaction, providing instructors with timely diagnostic metrics to enhance decision making during course implementation.

4. Develop more strategic collaborative, problem-solving tasks and exercises to increase participation and improve content, interpersonal, and intellectual interaction.

CONCLUSION

This article presented pilot study findings on using weblog facilities to increase participation and improve interaction in Web-based learning communities. The study revealed a number of aspects to assist educators in providing more structured and comprehensive online learning environments. The most important of these are people-related functions such as creating strategically-focused collaborative learning environ-
ments and improving the quality and quantity of, and accessibility to, course metrics to help educators make informed, timely decisions in online instructional environments. Accordingly, it is anticipated that the recommended strategies will improve participation and interactivity and contribute to more productive, supportive learning communities.

REFERENCES


KEY TERMS

Constructivism: Learning as interpretive, recursive, building processes by active learners interacting with physical and social worlds.

Information and Communication Technology (ICT): Refers to an emerging class of technologies—telephony, cable, satellite, and digital technologies such as computers, information networks, and software—that act as the building blocks of the networked world.

Learning Communities: Characterized as associated groups of learners, sharing common values, and a common understanding of purpose.

Portal: Considered an all-in-one Web site that acts as a centralized entranceway to all relevant content and applications on the Web.

“Sticky”: Web Features include chat rooms, e-mail links, and calendar functions.

Student-Centred Learning: Focuses on the needs of the students rather than on teachers and administrators, resulting in implications for the design of curriculum, course content, and interactivity of courses.

Weblog: A personal dated log format that is updated on a frequent basis with new information written by the site owner, gleaned from Web site sources, or contributed to by approved users.

This work was previously published in the Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 58-64, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 3.21
Innovative Technologies for Education and Learning: Education and Knowledge-Oriented Applications of Blogs, Wikis, Podcasts, and More

Jeffrey Hsu
Fairleigh Dickinson University, USA

ABSTRACT

A number of new communications technologies have emerged in recent years that have been largely regarded and intended for personal and recreational use. However, these “conversational technologies” and “constructivist learning tools,” coupled with the power and reach of the Internet, have made them viable choices for both educational learning and knowledge-oriented applications. The technologies given attention in this article include instant messaging (IM), Weblogs (blogs), wikis, and podcasts. A discussion of the technologies and uses, underlying educational and cognitive psychology theories, and also applications for education and the management of knowledge, are examined in detail. The implications for education, as well as areas for future research are also explored.

INTRODUCTION

For many years, the mediums employed for education have remained fairly constant and traditional: tried and true methods such as the blackboard and chalk, whiteboards, flipcharts, and overhead projectors. The employment of computing technologies has resulted in the use of PowerPoint, e-mail, and Web-based course portals/enhancements such as Blackboard and WebCT.

There have been numerous studies done, and papers written, about the use of technology in the classroom, together with work on the related areas of e-learning, Web-based learning, and online learning. The usage of computing technologies in education has been examined in numerous studies, and there is a sizable body of work on Web and online learning, including the studies by Ahn, Han, and Han (2005), Liu and Chen...
(2005), Beck, Kung, Park, and Yang (2004), and numerous others.

In particular, some of these technologies have been recognized as useful in the classroom, and have been engaged in innovative ways. The technologies of particular interest are those that are referred to as “conversational technologies,” which allow for the creation and sharing of information (KPMG, 2003; Wagner, 2004). Another term often used to describe these technologies is the concept of “constructivist learning tools,” which encourage, and are focused on, users creating, or constructing, their own content (Seitzinger, 2006).

The interest in employing these kinds of technologies stems not only from the unique pedagogical benefits gained, but also from the basic need to stay in tune with the focus and strengths of today’s students. Prensky (2001) suggests that the students being taught today are “no longer the people our educational system was designed to teach” and that while the students of today can be termed “digital natives,” many educators could be better termed “digital immigrants.” Yet another way to look at this is to view earlier educational approaches as “print-based,” while those of the current environment can be called “digitally-based, secondly-oral” (Ferris & Wilder, 2006).

The purpose of this article is to examine these technologies and explore both the evolution of their use from personal applications to that of educational tools, and also to examine the key educational applications for which these are being used. Relevant research and applications are examined and analyzed. The future of these technologies for educational and professional use, together with viable research areas, is examined as well.

CONVERSATIONAL TECHNOLOGIES AND CONSTRUCTIVIST LEARNING TOOLS

The notion of conversational technologies is not a new one, as it encompasses many types of systems that have been widely used for some time, including e-mail, video conferencing, and discussion forums.

The term “conversational technology” is derived from the work of Locke et al. (2000) relating to conversational exchanges and his Cluetrain Manifesto. One of the key concepts here is that “markets are conversations” and that knowledge is created and shared using question and answer dialog. Specific theses that relate to this form of “conversational knowledge management” suggest that aggregation and abstraction of information helps to create information. Other characteristics of conversational knowledge management include the fact that it is fast, stored in different locations, and does not require sophisticated technologies in order to be accomplished (Wagner, 2004).

Conversational technologies encompass a wide range of systems and software, many of which are familiar, including e-mail, instant messaging, Web pages, discussion forums, video and audio content/streaming, wikis, and Weblogs. While there are specific aspects that are of interest in terms of the more mature technologies, the ones that will be given attention in this article are the issues, impacts, and applications relating to IM, blogs, wikis, and podcasts. These are technologies that are newer, have a growing base of users, and are starting to become recognized as viable tools for education.

The term “constructivist learning tool” has also become associated with these, particularly blogs and wikis, in that they have a key characteristic of allowing users to develop and maintain their own content. Some of the characteristics of con-
Innovative Technologies for Education and Learning

Structivist learning include engagement, active learning, collaboration, real world based, and the usage of reflection as a part of the learning process (Seitzinger, 2006).

It should be noted that these technologies and tools are best suited to course structures where class collaboration and communication are encouraged, rather than those with an emphasis on lectures and a presentation of factual information. In addition, in courses where there is substantial group work, or projects where a collaborative document is created, the use of these would be especially helpful and useful. Both hybrid and full distance learning courses would be situations where these could also be used effectively.

TEACHING AND LEARNING: NEW TRENDS

Conversational and constructivist technologies are certainly here to stay, as evidenced by their extensive role in our society. It would therefore be useful to examine their applicability in the educational realm. While usage based on popularity or student preference seems to be one factor, there are also theoretical and conceptual bases for employing these kinds of technologies in the classroom.

Earlier paradigms of teaching emphasized print-based materials for instruction, which included printed textbooks, paper-based instructional materials, and written tutorials, all of which are grounded in the notion that the teacher, lecture, and instructional materials form not only the basis, but also the authority in the educational process. The transmission of material from the teacher (lecture) and/or textbook to the student (called the “print model”) is still the central basis of most teaching, even if they are supplemented with other methods including discussion and other forms of student interaction/participation (Ferris & Wilder, 2006).

However, the advent of digital and conversational technologies has brought forth the new concept of secondary orality (Ong, 1982). This concept emphasizes that teaching and learning should go beyond printed materials toward a greater emphasis on group work, fostering student communities, and encouraging student participation. The concept encourages a greater sense of interaction with and “ownership” of knowledge, emphasizing self-awareness and expression, and effectively using electronic tools (Gronbeck, Farrell, & Soukup, 1991).

The use of conversational technologies can have a positive impact, because they attempt to not only improve upon the print approach, but also use secondary-oral techniques. In other words, while a student can still be presented with material (in different formats) using the print model, the introduction of secondary-oral methods can be used to improve the overall learning experience. Using the latter, there is the opportunity to work and learn collaboratively, explore, analyze, engage in discussion, and otherwise “learn” in new and innovative ways (Ferris & Wilder, 2006; Wallace, 2005).

INSTANT MESSAGING (IM)

It is unlikely that there would be many college students who are unfamiliar with the use of IM. Allowing for interactive and real-time synchronous communications with instant response, instant messenger is truly conversational in that it allows for “chat” and communications between both individuals and groups. The major instant messaging systems in use include AOL (AIM), MSN Messenger, Yahoo! Messenger, and ICQ.

IM is a means for users to “chat” and communicate in real-time. While originally the domain of personal users, over time the unique benefits and effectiveness of this medium were realized, and IM started to become accepted as a form
of communication in businesses (particularly high-tech firms), and now has been studied and tested as an educational tool (Kinzie, Whitaker, & Hofer, 2005).

The important features of IM include both its synchronous nature and its ability to support both chat and phone-like interaction. While real-time interaction allows for rapid communications to occur, there is also no need to enter an interaction “space” as with chat rooms. Instead, the main usage of IM is in one-on-one communications, which can be more formally termed as a dyadic “call” model, which more closely resembles phone call interaction. It should be noted that even though

Table 1.

<table>
<thead>
<tr>
<th>INSTANT MESSAGING</th>
<th>Description</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Educational applications</th>
<th>Course/subject suitability</th>
<th>Theoretical foundations</th>
</tr>
</thead>
</table>
|                   | Real-time communications that allow for informal communications to be conducted easily and quickly | Availability and acceptance by students  
Social presence (know the status of other users online)  
Real-time (synchronous) communications  
Encourages collaboration  
Reduces formality in communications | Distracted attention, especially in a classroom setting  
“Time waster” that is not directed toward course content, but on personal discussions  
Expectations of 24-7 instructor access  
Can be time consuming for instructors  
Benefits are uncertain in classroom settings | Virtual office hours (instructor-student)  
Collaboration on group projects  
Real-time class discussions  
Mentoring | Courses with group projects and assignments  
Distance learning support | Active learning  
Dual (verbal and visual) processing |
much of the communication is done between two individuals, there are some systems that support multiparty instant messaging.

Some of the salient features of IM include the ability for users to see user details as to current status (online, idle, away, out to lunch), and also on a user’s changes in status (active, logged out, etc.). Lists of users can be displayed on the screen, so that contact can be made when desired. If a “chat” is initiated, a special window comes up, and the interaction can commence, provided that both parties are online and willing to proceed.

The real-time nature of IM has resulted in the technology being used for reasons aside from personal “chat.” In business, IM has become in some industries an accepted form of communication. A number of studies have concluded that instant messaging is ideal for informal interaction. In particular, the use of IM has been shown to be helpful in cases where collaborative coordination and problem solving is involved. Social bonding and interaction, which is a component contributing to the success of more complex collaboration situations, is also enhanced by using instant messenger technology (Nardi & Bradner, 2000).

An important difference between IM and e-mail is the tendency for instant messenger interaction to be more casual and informal than e-mails, which helps to bring about a more “friendly” communication atmosphere. This may in part be due to a reduction in the formalities that are typically involved when using e-mail or the phone. In particular, IM has been considered more suitable for such tasks as scheduling meetings, asking or answering quick questions, and for other kinds of tasks that are brief, require a prompt response, or are less formal. It is perceived to be far simpler to IM someone to ask a quick question, for example, or to confirm a meeting or lunch, rather than to e-mail or call (Nardi & Bradner, 2000).

It is also of interest that IM communications tend to be more flexible in terms of their uses (everything from task-related questions to a new joke), and can allow for greater expressiveness in terms of emotion, humor, and personality (Nardi & Bradner, 2000). Another interesting aspect is what Nardi and Bradner (2000) refer to as “outeraction,” which focuses on the processes associated with IM. These include conversational availability, communications zones, intermittent conversations, awareness, and conversational progress/media switching. IM is useful in certain communications situations, since it tends to be less disruptive and interrupting, while at the same time a user’s availability is more clearly known (scanning buddy list status, for example). It is also a convenient means for setting up more formal interactions, such as arranging a conference call (media switching). Intermittent, dispersed communications can be conducted over a longer period of time, which includes interruptions. Another benefit includes the knowledge that others are “there” and available, even if not currently in chat mode; however there is always the opportunity to make contact, whether through IM or a different form of communications.

While some educators may scoff at and even express criticism at the thought of instant messaging as a viable educational tool, others believe there is potential in the medium.

In terms of educational uses for IM, they are being explored and tested. Clearly, IM not only allows students to collaborate more effectively on homework assignments and projects, but also helps to maintain a closer social network between students, which could have a positive impact on learning. In addition, if IM is carefully targeted and focused toward the material or lecture topic in hand, the use of IM may actually help and stimulate deeper and more active learning.

On the other hand, it has been hypothesized that the distraction of working on various other tasks in parallel with IM, known as “distracted attention,” may have a negative impact on learning (Hembrooke & Gay, 2003).

Active learning (Grabinger, 1996) and dual (verbal and visual) processing (Clark & Paivio,
Innovative Technologies for Education and Learning

1991) are at work here. It could be said that using IM to encourage greater discussion and reflection on the course contents would be likened to the use of discussion boards; however, since IM is a real-time technology, the interaction is conducted during the lecture or class, not afterward. Some studies have reported positive effects and student satisfaction from IM being used to discuss course subjects in real-time (Guernsey, 2003).

A study by Kinzie et al. (2005) examined the use of IM during classroom lectures and found that while the general idea of using IM online discussions was positively received, the actual process and experience of using IM to conduct discussions during class lecture sessions was not found to be less than a positive experience by both teachers and students. It was suggested that the difficulties of multitasking and dividing one’s attention between the lecture and instructor, doing the IM discussion, contributed to the lack of satisfaction with the process.

Burke (2004) used instant messaging as a medium for creating course diaries in three different mathematics courses. IM was chosen since it was thought to be popular, widely used by students, and considered more “fun,” so there was some hope that this popularity would transfer over to greater and more enthusiastic usage by students. In fact, the choice was made to use IM over a seemingly more suitable choice, blogs. A bot was created that would retrieve IM diary entries from students and store them in a PostgreSQL database, and there was also a program set up to display diary entries from each student, viewable by both the student and the instructor. The main finding of the study was that the IM media was not ideally suited for all kinds of courses, especially those that involved creating longer portions of text, or involved diagramming. Error detection and recovery was also not that well developed, and also there was a need for better editing tools.

In summary, while instant messenger can be appropriate for various applications, in particular for information communications in a business setting, the results from educational studies appear to be mixed, with both positive and negative effects noted. While there seem to be advantages to real-time communications between students, between students and instructors, and also between groups working on a project, it appears that there are problems and limitations if the technology is used in a classroom setting. The challenges of focusing on a class lecture, together with maintaining a conversation online, seem to be a problem that has not yet been resolved. In addition, while instructors can often establish closer relationships with students using IM, there is also the problem of unreasonable student expectations of continuous teacher access, which may not be present if IM was not available as an option. In connection with this, using IM for student help can result in a greater time commitment, since sessions can become lengthy with many questions and responses being sent back and forth.

BLOGS (WEBLOGS)

Blogs started as a means for expressive individuals to post online diaries of themselves. Complete with text and photos, these logs were essentially an individual’s online narrative or diary, with events, stories, and opinions. While its original use was for personal expression, recently its effectiveness as a tool for education has been discovered, including its use as an extension of “learning logs,” which are created online (Barger, 1997). One of the earliest blogs, as we know and use them today, was Dave Winer’s Scripting News, which was put online in 1997. While the use of Weblogs can be considered generally new, the concepts of keeping a “log” or “learning log” is not.

The concept of “learning logs” has been in use since before the advent of the Weblog. The concept of this is to enable someone to document his or her learning, and also to do some critical reflection (Fulwiler, 1987) and self-analysis. The use of a learning log or journal is related to action research
learning strategies (Cherry, 1998) and attempts to link previous knowledge and new information learned. Blogs are a natural extension of learning logs/journals in that they are electronic and can be made available (“published”) more easily (Armstrong, Berry, & Lamshed, 2004).

The use of electronic Weblogs as educational tools offers the benefits of increased information sharing, simplified publication of information, and improved instructor monitoring and review (Flatley, 2005; Wagner, 2003). The use of blogs has been expanding, as Perseus Development reported that there were some 10 million blogs in 2004, and the number is ever increasing (Nussbaum, 2004). The growth in this area is expected to increase in the years to come.

Table 2.

<table>
<thead>
<tr>
<th>WEBLOGS (BLOGS)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>A technology that allows a sequence of entries (online diary, journal) to be posted and published online</td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td>Reflection and critical thinking are encouraged</td>
</tr>
<tr>
<td></td>
<td>Authenticity through publication</td>
</tr>
<tr>
<td></td>
<td>Social presence</td>
</tr>
<tr>
<td></td>
<td>Development of a learning community</td>
</tr>
<tr>
<td></td>
<td>Active learning encouraged</td>
</tr>
<tr>
<td></td>
<td>Ability to receive and respond to feedback</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>Controlled primarily by blog author</td>
</tr>
<tr>
<td></td>
<td>Editing/modifications not open as in a wiki</td>
</tr>
<tr>
<td><strong>Educational applications</strong></td>
<td>Online learning journal</td>
</tr>
<tr>
<td></td>
<td>Problem solving/manipulation space</td>
</tr>
<tr>
<td></td>
<td>Online gallery space (writings, portfolio, other work)</td>
</tr>
<tr>
<td></td>
<td>Peer review exercises</td>
</tr>
<tr>
<td><strong>Course/subject suitability</strong></td>
<td>Writing courses</td>
</tr>
<tr>
<td></td>
<td>Foreign language courses</td>
</tr>
<tr>
<td></td>
<td>Research seminars</td>
</tr>
<tr>
<td><strong>Theoretical foundations</strong></td>
<td>Activity theory</td>
</tr>
<tr>
<td></td>
<td>Guided discovery</td>
</tr>
<tr>
<td></td>
<td>Cognitive scaffolding</td>
</tr>
<tr>
<td></td>
<td>Receptive learning</td>
</tr>
<tr>
<td></td>
<td>Social cognition</td>
</tr>
<tr>
<td></td>
<td>Community practice</td>
</tr>
<tr>
<td></td>
<td>Communities of inquiry</td>
</tr>
</tbody>
</table>
Blogs can be defined more formally as being “frequently updated Web sites consisting of dated entries arranged in reverse chronological order” (Walker, 2005) and can take several forms, including the personal diary/journal, knowledge-based logs, and filter blogs. The first, an electronic, online diary of one’s life events and opinions, is probably the most common. The online diary/journal blog is one that, being on the Internet, is public, as opposed to the traditional (typically paper) diaries, which are generally kept private. It should come as no surprise that there are many different online diary/journal blogs that are currently online, where one can find out details, often much more than one might want to know, about someone’s life and thoughts. Personal blogs form the majority of the blogs that are currently online and available, which make up roughly 70% of all the blogs in existence (Herring et al., 2003).

The second type (knowledge-based) captures knowledge and places it online in various formats. The third type (filter) attempts to select, rate, or comment on information contained in other sites (Herring et al., 2004).

There are software packages that are designed to help users create blogs, including Blogger, Xanga, Blurt, and MovableType. While the basic features of most blog software emphasize the creation of blog pages, some of the more sophisticated ones offer the capability to track readership, see who followed what links, add photos, and set up more advanced structures. When online, blogs can range from being totally public (listed in the blog service directory), to being “unlisted” but still open to being highly restricted (password-protected).

Blogs are also interesting and unique in that they are not merely online versions of paper diaries and journals. Rather, as a communications medium under the control of the main writer (author), it is reflective of the fact that an audience is “watching and listening.” What is put on a blog is not merely a one-sided set of thoughts and reporting of events; there can also be responses to feedback and reactions from the “viewing audience.” Therefore, blogging is considered a highly social activity, rather than a personal one.

In fact, recent work has indicated that the momentum for creating, and also updating a blog, came about as a result of encouragement from friends and the viewing audience (Nardi et al., 2004). In addition, blogging can have negative repercussions when posted information is perceived to be confidential, proprietary, or improper. In some cases, employees posting what was considered by their employers as “confidential” information can cause problems.

Blogs do not, in general, exhibit a free-flow of information between the blogger and the outside audience. While feedback is often requested, received, and desired by the blogger, the level and quantity of feedback from readers is generally limited compared with the output from the blog writer. In addition, while blogs may have sections where hyperlinks are mentioned, the number of hyperlinks in blog pages is frequently not very large (Schiano, Nardi, Gumbrecht, & Swartz, 2004).

More formally, Weblogs can be considered to be objects motivating human behavior, which is related to activity theory. Activity theory states that there are objects that have motives that respond to a human need or desire, and that they manifest a person’s desire to accomplish that motive (Leontiev, 1978; Vygotsky, 1986). The objects that connect bloggers to their own social networks include letting people know what is happening in their lives, voicing opinions, asking for feedback, and “letting off steam” about certain challenges or difficulties currently being experienced, to name a few (Nardi et al., 2004).

Blogs have been categorized by Krisnamurthy (2002) as being categorized into four different types, along the dimensions of individual versus community, and personal versus topical. A blog can therefore range from being very individual...
Innovative Technologies for Education and Learning

and personal, all the way to being open to the community, however very focused on a particular topic.

The acceptance of blogs for educational purposes is gaining interest, with one university, the University of Maryland, attempting to implement blogging software campus-wide (Higgins, Reeves, & Byrd, 2004).

In addition, the educational uses of blogs take advantage of their ability to encourage expression and the development of online relationships. Blogs allow for learning and interaction to be more knowledge-centered, especially if the assignments are structured in the format of encouraging feedback and input from the instructor and outside experts. Blogs also allow students to gain a better understanding of a subject’s knowledge domain (Glogoff, 2005). As an example of this type of blog-enhanced class structure, students might be provided with a Weblog from which to complete certain course assignments. After researching the indicated subject, the student would present the new information by “publishing” it to the Weblog. The Weblog would constitute the student’s assignment, which would then be subject to review and critique by not only the instructor, but also by other students in the class. Supplementing this could be discussion boards, where threads would be devoted to discussions on the Weblogs created by the students. This kind of assignment and interaction would be especially useful for both hybrid and fully online distance learning courses (Glogoff, 2005).

There are other benefits of Weblogs. These could be expressed using the learning theories and concepts of guided discovery, directive learning, receptive learning, and social/community-centered instruction.

Guided discovery allows for the exploration and study of a certain topic, which is then followed by assignments that emphasize the synthesis of information. In effect, a student can be asked to research an area and “construct knowledge” using the Weblog as a medium. Part of the assignment goes beyond merely explaining or presenting the material, and asks for the application of the concept using a real-world situation. The ability for students to post and make comments about other students’ blogs provides an atmosphere of interactivity and collaboration. One of the advantages of using blogs together with guided discovery is that it encourages the use of cognitive scaffolding, where students would approach learning (together with the use of blogs and interaction) by repeatedly seeking information, reflecting and thinking about what has been learned, and then going back and obtaining more information, so as to build upon and dig deeper into the subject area. This can result in a more active and productive form of learning (Betts & Glogoff, 2004; Glogoff, 2005).

Directive learning, where responses from students are followed by prompt feedback from instructors, can also be supported using blogs. In this case, students would not only use a blog to submit assignments, but also to review instructor feedback. In addition to feedback, there would be opportunities for the instructor to ask additional questions, in effect, to encourage further exploration and “drilling down” into the subject (Betts & Glogoff, 2004; Glogoff, 2005).

Receptive learning is where instructional modules are presented that focus on certain broader areas, from which certain sub-areas within these are highlighted for a student to research and report on. Generally, the areas are contained within a designated theoretical context (Betts & Glogoff, 2004; Glogoff, 2005).

Social/community-centered instruction is a logical component of educational work using blogs, and in particular the use of peer and social interaction as a part of the learning process. The use of blogs functions as an easily accessible medium for students to present their findings (and to be read by others) and also to integrate not only the information presented, but also related links and references to other resources. This form of
interaction helps to encourage further exploration by students. A blog-based discussion can then be continued by conducting peer reviews of other students’ blogs, which may include commentary, critique, the posing of questions, and opening up the way for further inquiry. The ability to share and benefit from the findings of other students (and to explore further) is another important outcome. The theories of community practice (Snyder, 2002), social cognition (Vygotsky, 1978), and communities of inquiry (Lipman, 1991) provide support for the blog-related techniques mentioned previously.

Ducate and Lomicka (2005) discuss their experiences in using Weblogs to support foreign language classes. Weblogs help the foreign language student to learn during the process of reading, and then creating blog entries. Students can learn by reading blogs that are written in the new, target language, including learning new vocabulary, checking out links and further information on words, and learning associated cultural information. The reading and absorption of blogs on the culture associated with the target language, including literature and lifestyles, all would contribute to the learning process.

Another approach would be to have students maintain blogs written in their new, target language, and then the goal would be to seek commentary and critique on these blogs by others in the class. Yet another innovative method might be to share blogs with other classes studying the same language, and for students to read and comment on each other’s postings. In the case where students travel to a country where the target language is spoken, the compilation of travel blogs would be a useful learning tool as well (Ducate & Lomicka, 2005).

Wang and Fang (2005) looked at whether the use of blogs encouraged or enhanced cooperative learning in an English rhetoric/writing class taught in Taiwan. The main premise was that blogs can encourage students to spend more time working within a class “community” and can benefit from a greater sharing of contributions and inputs. In general, cooperative learning benefits can be divided into three different types: formal, informal, and base groups. Formal cooperative learning is where the instructor explicitly provides course materials and assignments to a group and then observes the students’ learning processes. When the instructor provides information more generally (such as detailing how to use a blog for course assignments) and then lets the group work out their own methods for handling an assignment, that is known as informal cooperative learning. When a learning-oriented group is maintained for an extended period of time, such as throughout a semester, then this form of cooperative learning is known as a cooperative base group (Johnson & Johnson, 1998; Johnson, Johnson, & Holubec, 1991). The study, run over the course of a semester, found that the use of blogs contributed not only to cooperative learning in general, but also to autonomous learning. Autonomous learning is focused on how much students take responsibility for their own learning, and also develop self-confidence in the task or skill (Wenden, 1991). The use of blog technologies was also found to help improve information processing, learning self-evaluation, and effective time management (Wang & Fang, 2005).

Martindale and Wiley (2005) also used blogs in their courses and looked at two cases of the impact of this technology on teaching and learning. Martindale taught a doctoral-level course on instructional design and technology. In it, students were introduced to blogs and used them throughout the course, which overall tended to promote higher levels of quality in their course work. Blogs were used to post ideas and abstracts of their projects, and also to place links for relevant research papers and Web-based resources. The end result was a course “knowledge base” that represented the cumulative output of the students in the course. The blogs were also used for article critiques,
which were an integral part of each weekly class. Students were given access to the blogs of other students and were able to offer feedback.

Wiley taught two different courses, one on online learning research, and the other on online interaction culture. Both included the use of blogs as a supporting technology, the first employing a general course blog where information about the course, student assignments, and class updates and student/instructor interaction exchanges were posted on an ongoing basis. In the second, blogs were used to discuss experiences using different online communications technologies, causing students to become highly engaged, resulting in passionate discussions and detailed commentaries posted to the blogs, far exceeding the level and depth of feedback that was expected (Martindale & Wiley, 2005).

In summary, blogs can be useful for educational purposes, particularly where there is the need to encourage and stimulate critical thinking and reflection on a work, concept, or idea. The submission or publication of a document or text as a blog can then lead others in a class to review and comment, setting the stage for greater analysis and study. In particular, blogs are suited to writing courses, where a text can be analyzed and critiqued, or for a situation where a peer review is desired. The blog approach is also considered useful for group study of a certain focused problem or case, such as those used in graduate courses and seminars.

WIKIS

Yet another technology, known as the wiki, has emerged, which allows for improved collaboration compared with Weblogs. While the major emphasis of Weblogs is the creation of a set of pages and documents primarily by a single individual, the strength of a wiki is the ability for numerous interested readers and users to express ideas online, edit someone else’s work, send and receive ideas, and post links to related resources and sites. As a result, wikis go a step further and allow for greater collaboration and interactivity (Chawner & Gorman, 2002). Wikis have been found to have value for educational purposes, and their use has begun to be integrated into a number of university courses, in particular (Kinzie, 2005).

The term “wiki” comes from the Hawaiian word “wikiwiki,” which means “fast.” The technology is computer-based and can be generally described as a knowledge sharing and creation system that has as its basis a set of Web pages, which can be created and updated on an iterative and collaborative basis, and is in many ways a form of groupware. A wiki is designed to run on the Internet and World Wide Web, uses the HTTP protocol, and resembles traditional Web sites in terms of its underlying structure. Some of the benefits of wikis include the ability to easily create pages (using a simplified form of HTML or basic HTML) and the ability for a document to be authored collaboratively and collectively. In particular, simplicity is the key to wikis, and wiki pages have been designed to be easy to create, (simpler than the process of creating standard Web pages). One of the better-known examples of a wiki is www.wikipedia.org, which is an online encyclopedia with entries authored and edited by different persons worldwide, and in several different languages as well. In essence, it is an online information resource that is authored by interested and knowledgeable persons from around the world.

Wagner (2004) developed a set of design principles that relate to wikis. These are the principles of open, incremental, organic, mundane, universal, overt, unified, precise, tolerant, observable, and convergent wikis. Open means that anyone can edit a wiki, creating an “open
source” environment for the sharing of knowledge. *Incremental* means that new pages can be added, even if they do not yet exist. *Organic* means that the information can be continuously evolving, as changes and edits are made. Wikis are *mundane* because they involve the use of a simplified set of essential commands. The design of wikis is also *universal*, meaning that writing and editing is a “combined” activity, formatting is related to input (overt), page names are not context specific (unified), and pages are generally named with some precision (precise). Wikis should be *tolerant* of error; activity should be *observable* by all, and duplications are undesirable and should be deleted (convergent).

There are a number of software programs that enable the effective creation of wiki pages, including TikiWiki, TWiki, and Pmwiki. These allow for the effective creation, modification/editing, and management of wikis, including creating pages, creating links, formatting, and feature modules (discussion forums, photo pages, download areas, etc.) (Chawner & Lewis, 2006)

Wikis are set up to allow for easy collaboration, and more specifically, editing. Rather than passively reading a passage of text or related information (which may include graphics, multimedia, hyperlinks, etc.), a reader of a wiki can also take on the role of a writer, making changes to the text (re-organizing, editing, re-writing, and marking up) at will. In essence, the document is open to changes by a “collaborative community,” which allows for the secondary-oral model in education to be applied.

**Table 3.**

<table>
<thead>
<tr>
<th>WIKI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>A technology that allows for material to be easily published online, and also allows open editing and inputs by a group</td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td>Contributions and editing by a group</td>
</tr>
<tr>
<td></td>
<td>Open access to all users</td>
</tr>
<tr>
<td></td>
<td>Collaborative</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>Lack of organization and structure may result in an unmanageable wiki</td>
</tr>
<tr>
<td></td>
<td>Tracking of contributions and modifications can be difficult</td>
</tr>
<tr>
<td></td>
<td>Quality control</td>
</tr>
<tr>
<td><strong>Educational applications</strong></td>
<td>Collaborative writing/authoring</td>
</tr>
<tr>
<td></td>
<td>Group project management</td>
</tr>
<tr>
<td></td>
<td>Brainstorming activities</td>
</tr>
<tr>
<td></td>
<td>Knowledge base creation (knowledge management)</td>
</tr>
<tr>
<td><strong>Course/subject suitability</strong></td>
<td>Knowledge management</td>
</tr>
<tr>
<td></td>
<td>Writing</td>
</tr>
<tr>
<td></td>
<td>Group work in courses</td>
</tr>
<tr>
<td><strong>Theoretical foundations</strong></td>
<td>Conversational technology</td>
</tr>
<tr>
<td></td>
<td>Constructivist learning tool</td>
</tr>
</tbody>
</table>
One reservation on the part of educators to embrace wikis is the fact that wikis are designed to allow for open and free access and editing by all members of a “community.” As a result, if improperly managed, a wiki’s content can become an unreliable, inaccurate, or biased source of information due to its largely unmonitored format. There is also the issue of having little or no “quality control,” resulting in a wiki not being trusted by its readers and users. An example of this was the controversy over the accuracy and reliability of Wikipedia, in the case of John Seigenthaler, in which the subject alleged that false and incorrect statements were posted in his biography (Seigenthaler, 2005). However, other studies have attempted to prove that Wikipedia was overall an accurate information source. One article reported that after analyzing a set of Wikipedia’s science articles, they were judged to be, by the British journal Nature, as reliable as the Encyclopedia Britannica (BBC News, 2005).

Wikis are useful for education in that they help to promote student participation and also a sense of group community and purpose in learning. Indeed, an important element of this is the relaxed sense of control over the content, allowing students to have a greater role in managing its focus and direction.

Wikis are not all the same, and there is significant diversity between various forms and implementations of wiki systems. In fact, it could be debated what features truly characterize a “true” wiki. The features inherent in most include the ability for users to both read and edit information, without the need for security or access restrictions. The emphasis is on simplicity, and the informal, “never finished” nature of wikis, which may constitute the contributions of multiple authors, is another key characteristic. While the emphasis of many wikis is on simplicity and a lack of access restrictions, that does not mean that all wikis work this way. In reality, there can be a continuum of features from simple to complex. At the complex end of the scale can be capabilities for security/access restrictions, complex organizational structures, and for integrated with content management systems (Lamb, 2004).

Now that the strengths and weaknesses of wikis have been established, it would be useful to examine the educational applications of wikis. In general, the most suitable applications are those that take advantage of the wiki’s free, open structure. As such, the use of wikis as discussion/bulletin boards, brainstorming tools, and online sketchpads is appropriate. Meeting planning is another viable application area, in that the organizer can start with a preliminary agenda, from which the other participants can then add their own additions or make modifications or comments.

An important application area for wikis has been identified in knowledge management (KM). The use of wikis for knowledge management may allow for an improvement over existing systems and technologies. Currently, with existing KM systems, there does exist a number of bottlenecks relating to knowledge acquisition, namely acquisition latency, narrow bandwidths, knowledge inaccuracy, and “maintenance traps.” Basically, these knowledge acquisition bottlenecks result from a time lag between when the knowledge is created, and then distributed. In addition, there are the problems of limited channels of knowledge input, possibilities of erroneous information being received, and also the difficulties of maintaining the knowledge base as it grows larger (Land, 2002; Wagner, 2006; Waterman, 1986).

The use of wikis to elicit a “bazaar” approach to knowledge management, rather than a “cathedral” approach, is proposed as a better alternative. These terms are derived from software development, whether the “bazaar” allows for more continuous and open access to code (or information), as opposed to the “cathedral” approach where access is only provided on certain (release) dates to certain persons. The difference between the “cathedral”
(closed), sources of knowledge acquisition management and “bazaar” (open) could be illustrated by the difference between encyclopedias that are created by a single firm, such as Encarta or the Encyclopedia Brittanica, and those that obtain information from readers and users, such as the well-known Wikipedia.

The emphasis therefore is on teamwork, continuous review and testing, and the development of conversational sharing (Wagner, 2006). Inherent in the workings of wikis is support for an open, collaborative environment, where many people can contribute to the development of knowledge instead of being limited to a set of “experts.” It appears that conversational knowledge acquisition and management are appropriate for wikis (Cheung, Lee, Ip, & Wagner, 2005). As for educational applications and KM, a study by Raman, Ryan, and Olfman (2005) examined the use of a wiki to help encourage and support collaborative activities in a knowledge management course. More specifically, using wikis in the course helped to encourage openness and better sharing and updating of knowledge bases. Many-to-many communication is supported, and the persistence of the created pages formed the basis of a knowledge repository. In short, the impact of easy page creation and improved updating and editing, together with effective maintenance of knowledge histories, were seen as positives (Raman et al., 2005; Bergin, 2002).

Activities in the KM course activities included group article review assignments, answering questions about sharing knowledge and uses of the wiki technology, and also creating a wiki-based knowledge management system. Students were asked to create, update, refine, and then maintain a class knowledge management system. In terms of these experiences, while the use of the wiki technology was generally viewed positively, feedback received indicated that, since the goals of using the wiki were not made clear, using one was perceived to be counter-productive. More specific guidance on goals and objectives, a clearer system structure, and advanced training were suggested as ways to make the wiki a more effective educational tool. The availability of too many features made the task of doing the course activities more difficult, since much time was spent learning the various features rather than focusing on the task at hand. A simpler, less feature-rich version was therefore preferred (Raman et al., 2005).

Another popular application of wikis in the classroom is in supporting writing courses. The use of this technology can help to foster the impression that writing is “fun,” and that there can be a shared and collaborative side to writing, revising, and commenting on written work. In other words, the technology can benefit not only the writing and editing process, but also in bringing about an awareness that writing is being done for a specific audience.

An example of the use of wikis in English is the Romantic Audience Program at Bowdoin College, where students used a wiki to discuss and examine Romantic literature, focusing on poetry, poets, and related topics. The technology was used to elicit discussion and analyses by the group, encourage elaboration on areas where further study or insight was sought, and to seek out linkages to additional sources and commentary. Another project is Auburn University’s Wikifish, which was created by one school within the university, where questions are posed, and opinions and comments by all are encouraged.

Difficulties encountered in using wikis for education include the difficulty of tracking the new pages and contributions made by students, since modification can be made to another student’s text without specifically identifying the author or editor. As a result, it can be difficult to monitor, and properly attribute, what contributions were made by whom, on a particular page. A proposed solution to this was an instructor’s use of categories and topics, requiring that students link to and add, rather than simply modify, the contributions of other students. Another issue was how much of a balance in terms of the tradeoff between
Innovative Technologies for Education and Learning

total freedom, and total control, was ideal. Since
a clear benefit of a wiki is the emphasis on free
expression and on spontaneous inputs, reducing
this may have a negative effect on open interaction
and student contributions (Lamb, 2004).

An interesting application of the use of wikis
in the classroom was the work by Bruns and
Humphreys (2005), where wikis were used in a
New Media Technologies course to collaboratively
develop entries for an online wiki encyclopedia
called the M/Cyclopedia of New Media, an ac-
tual live wiki resource made available online.
The effort to develop entries involved over 150
students spanning six classes in the program.
Feedback indicated that while students had little
difficulty with using the wiki system, obstacles
came about more with the writing and format of
the encyclopedia entries.

Another study examined the use of wiki
technology in German-speaking communities
in Switzerland (Honegger, 2005), while Desi-
lets, Paquet, and Vinson (2005) looked at how usable wikis are and found that the main prob-
lems encountered related to the management of
hyperlinks. Wikis were also examined from the
perspective of structure: how did the use of cer-
tain templates affect the appearance and layout
of wiki pages? The results suggested that they
are useful in certain circumstances and could be
helpful overall to end users (Haake, Lukosch, &
Schummer, 2005).

In summary, wikis are best suited to course
and activities where there is a document, text, or
other project to be worked on jointly by a class
or group. In a sense, it is a tool for collaboration,
and a form of groupware. The compilation of a
class or group report or project, the creation of a
knowledge base, or brainstorming sessions ap-
pear to be viable applications. The free and open
structure, however, can fall prey to disorganization
and degradation in quality, and so it is important
to have safeguards and procedures in place to
ensure an effective result.

PODCASTS

While the terms “pod” and “podcast” at first
mention might evoke visions of “Invasion of the
Body Snatchers,” for most tech people in the
know, the reference to Pod is almost certainly a
reference to Apple’s popular and ubiquitous iPod.

However, podcasts are in actuality not what
their name might imply them to be. A “podcast,”
a combination of “iPod” and “broadcast,” neither
refers to a technology specifically requiring an
iPod, nor broadcasts information to users. Instead,
podcasts are multimedia files (typically audio or
video) that are downloaded to users on a subscrip-
tion basis. Because of the potential confusion due
to the use of the word “pod,” some have called for
the letters to mean “personal option digital” or
“personal on demand,” rather than iPod.

Podcasts can be played back on any device or
system that can play digital audio (typically MP3)
or video files, and are not broadcast to a large au-
dience, in the way that television, radio, or spam
e-mails are sent. Instead, they are sent to users who
have specifically subscribed to a podcast service,
and as such, files are automatically downloaded
to the user’s computer when they are ready and
available. In addition, podcast files are generally
not streamed (as video is streamed), but rather
are downloaded for later playback (Lim, 2005;
Lum, 2006). Podcasts are delivered to subscrib-
ers through the use of RSS or RFD XML format
media feeds, rather than more traditional forms
of downloading (Descy, 2005).

Podcasts are considered to be a viable educa-
tional tool for several reasons. First, because of
the popularity and wide use of devices such as
the iPod and similar units, it would seem like a
good medium from which to distribute educational
materials. Secondly, the ease with which informa-
tion can be retrieved and accessed makes this
a good choice for students, who are using these
devices on a regular basis for music and should
have few technical difficulties or learning curves
(Lum, 2006).
There are multiple facets to podcasts. First, there is the consumption perspective, where someone downloads the podcast and then listens or views it. This involves subscribing to a service (or enrolling in a course), finding the relevant file, and then downloading and playing it. Alternatively, someone can actually create podcasts; an instructor can produce lessons for students, or students can produce an assignment in the form of a podcast file (Lum, 2006).

Education is one area where the power of the podcast has been used in various ways. At Duke University, orientation material was distributed as podcasts, loaded onto iPod units, and given to students in its 2004 incoming freshman class. The units were provided not only for orientation purposes, but also for use in playing podcast lectures when the students take certain courses at the university. At Mansfield University, students were sent podcasts about various student life issues, and at Arizona State University, President Michael Crow used podcasts to deliver messages to the university community (Lum, 2006).

While there appear to be sound reasons for using podcasts, there is also a theoretical basis behind the use of podcasting. This is based on cultural-historical activity theory (Engestrom, 2002) and is based on the fact that podcasting

### Table 4.

<table>
<thead>
<tr>
<th>PODCASTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Educational applications</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Course/subject suitability</strong></td>
</tr>
<tr>
<td><strong>Theoretical Foundations</strong></td>
</tr>
</tbody>
</table>
can be considered a tool that can be used to help learners to better interact with or understand a task and its environment. Vygotsky (1978) argues that the effectiveness of podcasts rests in its linkage between social/cultural influences present in the environment and the cognitive development of the learner. Expressed another way, the concept is that since so many student have access to iPods and MP3 players, it would make sense to explore the viability of using such a device for learning and educational purposes.

Lim (2005) discussed experiences in using podcasts for teaching geography. Because of the nature of the subject, it was found that video would have a greater impact and result than audio. Students were asked to submit assignments to be created and submitted as podcasts. Overall, this helped to bring about satisfaction and interest in terms of the subject. Ractham and Zhang (2006) looked at the potential for podcasts in a classroom setting. While there are the obvious benefits in terms of being able to distributing class materials, there also are benefits in terms of contributing to social networking in the class and continuing a flow of academic knowledge. Class discussions, announcements of research activities and conferences, and also campus activities could be distributed as podcasts. Review materials could be distributed effectively on an automatic basis to interested students. In addition, the ability for students to create podcasts to be distributed to others would also be a new means of submitting assignments or expressing creativity.

Podcasts, unlike IM, blogs, and wikis, offer a greater emphasis on providing engaging auditory and visual course materials to students, rather than on collaboration and group work. While not generally a substitute for traditional lectures and knowledge presentation, they offer the benefits of easily accessible and “digestible” course material. Whether it is excerpts from class lectures, highlights from a guest speaker, or an oral test review, the use of podcasts provides a means by which students can obtain and easily receive course-related information. In addition, it also provides a means by which students can express and present their work, which can then be “published” and distributed in podcast format.

DISCUSSION AND CONCLUSION

The face of education, whether online, hybrid, or classroom, is constantly changing, and it is important for educators to stay abreast of the many opportunities and possibilities that are available.

In this article, several technologies, generally termed as conversational technologies due to their interactive and collaborative nature, were discussed in detail, together with their capabilities, benefits, and educational applications. Relevant research and case studies as they relate to classroom and educational applications were discussed.

In general, the tools discussed here fall into the class known as “conversational technologies” or “constructivist learning tools.” As such, they emphasize student interaction, group learning, and collaboration, rather than the more traditional classroom mode. In light of this, they are more suited to educational courses or environments where the emphasis is on student communication, where students have access to technology, and where creative output and thinking is encouraged.

In the situation where a course is primarily lecture-based, or is mainly concerned with the delivery of factual or conceptual information, these tools may have limited applicability. The one application that may be helpful in this case may be for interaction to be extended outside of the classroom, through the use of instant messenger, or for supplemental materials to be distributed as podcasts.

Since each of the tools has its own characteristics and suitable applications, it would be up to the
Innovative Technologies for Education and Learning

The educator to select those that are most appropriate to one’s course and activity.

Instant messenger, which is commonly used by students for personal use, has found its way not only into the business community, but also into the classroom, because of its strengths in terms of informal communications that are conducted real-time. There are some mixed results regarding the use of IM for use for education; benefits are claimed by some, but there are limitations as well. The use of IM would best be employed in situations where group work, student communication, and real-time discussion would be helpful. However, it should be used cautiously, since it can be distracting, and students may end up carrying on personal, rather than course-related discussions.

Both blogs and wikis have been hailed as useful tools for education, and the specific advantages and disadvantages of each are noted and discussed. Blogs tend to be more one-sided, with an author presenting his or her information, with generally limited input from the readers and public. However, the use of the technology has been used effectively to promote information sharing and to support writing courses. The use of blogs to support online learning journals, class article/peer reviews, creating online portfolios and galleries, and for solving a specific problem or case would be advantageous. It would also appear to be a good medium for graduate research seminars, where papers and studies are analyzed and critiqued.

Wikis, which allow freedom in creation and in editing and enhancement by others, are especially useful in collaborative situations where the input of an entire group is desired instead of a single person’s contribution. In the classroom, wiki support for collaborative writing and group activities, where students contribute to the creation of a common result, would be useful. In general, any kind of learning activity that involves the collection and maintenance of knowledge or information may benefit from the use of wiki technology.

The use of podcasts, which may include audio or video, is growing in popularity and is being used for delivery of information to subscribers on an automatic basis. Educational podcasts, both for the delivery of audio and video-based knowledge to students, and also as a new medium for the creation and presentation of assignments, appear to have potential. While podcasts are useful as a means for publishing and distributing files of multimedia-based class materials, there also exists the potential to create new podcast content, both for educators and as a course activity.

Clearly, the use of these new conversational technologies is allowing for the enhancement and continued evolution of new and innovative forms of support for teaching and learning.

FUTURE RESEARCH AREAS

Certainly, there are many benefits to the use of conversational technologies and constructivist tools for educational use. However, more research needs to be done, not only in terms of identifying additional types of applications and uses, but also in terms of how to more effectively identify and apply new approaches to learning with the aid of these kinds of technologies.

Some of the broader research issues that can be examined include measuring both learning, and the perceived quality of education, depending on the specific technology or tool employed. Are there measurable benefits to using a certain technology in terms of the material learned, better class performance, or more subjective factors? It would also be useful to determine, particularly when using blogs and wikis, the neutrality or bias in the entries, and how much these contribute to (or detract from) the material submitted for a course assignment.

Other research areas are more technology specific. It was mentioned earlier that wikis can be a useful tool in knowledge management. The application of wikis to effective knowledge man-
Innovative Technologies for Education and Learning

agement deserves further attention, both in terms of developing or adapting the wiki structure and features to knowledge management uses, and also for identifying various kinds of user interfaces and functionality that would improve usability. The establishment of wiki-supported communities of practice is one area where the tool could possibly be useful.

Podcasting also has many areas that are ripe for further investigation. There are issues that can be explored in the areas of knowledge management, collaboration, and the adoption of podcasts. Some of the specific topics of interest include the management and sharing of knowledge using podcasts, examining whether their use actually improves learning, studying their effects on collaboration and networking, and what the factors (or features) are that would help to promote its use.

There also has been work on the psychological aspects of distance learning and online courses (Dickey, 2004), and a study of learners’ reactions to using IM, blogs, and wikis, for example, would yield insights into its appropriateness for its further use in education. Does the use of these technologies contribute to the satisfaction of students, or is more classroom face-to-face still better?

The realm of these new technologies is certainly ripe with a host of opportunities for both interesting and meaningful research studies.

REFERENCES


Cheung, K., Lee, F., Ip, R., & Wagner, C. (2005). The development of successful online communi-


Chapter 3.22
Ubiquitous Computing Technologies in Education

Gwo-Jen Hwang
National University of Tainan, Taiwan

Ting-Ting Wu
National University of Tainan, Taiwan

Yen-Jung Chen
National University of Tainan, Taiwan

ABSTRACT
The prosperous development of wireless communication and sensor technologies has attracted the attention of researchers from both computer and education fields. Various investigations have been made for applying the new technologies to education purposes, such that more active and adaptive learning activities can be conducted in the real world. Nowadays, ubiquitous learning (u-learning) has become a popular trend of education all over the world, and hence it is worth reviewing the potential issues concerning the use of u-computing technologies in education, which could be helpful to the researchers who are interested in the investigation of mobile and ubiquitous learning.

UBIQUITOUS LEARNING: THE NEW AGE FOR EDUCATION
In recent years, digitalization around the globe has been proceeding toward wireless communication and sensor technologies, which are able to detect the contexts of our daily lives, and provide personal supports accordingly. Such technology has been called ubiquitous computing (u-computing). The innovation and advance of those new technologies have led to a new research issue in education; that is, to develop an innovative learning environment so that the students can learn in any place at any time. Moreover, with the help of context-aware (sensor) technology, the learning system is able to detect the student learning behaviors in the real world, and hence
more active and adaptive learning activities can be conducted. Such a learning scenario is called context-aware ubiquitous learning (context-aware u-learning), which has gradually become a popular trend of education.

Researchers have demonstrated how a context-aware u-learning environment can be used to help the learners in increasing their ability for solving problems in the real world. For example, a group of researchers in Japan has employed u-computing technologies to conduct students to learn Japanese under real-world situations. The systems can provide learners with appropriate expressions according to different contexts (e.g., occasions or locations) via mobile devices (e.g., PDA, Personal Digital Assistant).

**Ubiquitous Computing Technologies that Facilitate Education**

U-computing technologies can be used to provide personalized services in the context-aware u-learning environment. For instance, when a student enters a lab or stands in front of an instrument, the context-aware devices are able to detect the location of the student and transfer the information to the server. Based on the decision of tutoring program in the server, relevant information, such as the operating procedure for each device, the need-to-know rules for working in the lab and emergency handling procedures, will be timely displayed to the student based on the personal and environmental contexts. Some ubiquitous computing technologies that might be useful in educational applications are given as follows:

- **Sensors for Detecting Personal Contexts**

  Researchers have proposed several significant characteristics of u-learning, which make it different from conventional e-learning, including seamless services, context-aware services, and adaptive services. In an ideal context-aware u-learning environment, the computing, communication and sensor equipment will be embedded and integrated into the articles for daily use. In addition, researchers also indicated that “time” and “location” might be the most important parameters for describing a learner’s context.

  There are several ways to detect the timely location of a learner. GPS (Global Positioning System) is one of the popular technologies for continuously detecting an object’s position by satellites, which trace air waves shot from the IC chips embedded in the objects. The object’s location is described with longitude, latitude and elevation. Other sensors, such as RFID (Radio Frequency Identification), which is an automatic identification method relying on storing and remotely retrieving data using devices called RFID tags or transponders, can also be used to detect the location of a learner by reading the messages from the tags, and then calculating the learner’s position based on the intensity of the signals.

- **Advanced Technologies for Detecting Personal Contexts**

  Learners might feel distressed or confused while encountering problems in the u-learning environment. Under such circumstances, a u-learning system could actively provide timely hints or assistance if the contexts concerning human emotions or attitudes can be sensed. Recent studies have depicted the possibilities for detecting such advanced personal contexts. Sensing devices with affective aware ability can not only capture the expressions of human faces, but also tell apart their emotional conditions. For example, the Affective Computing Group in MIT Media Lab of America have presented significant progress in this field, which can be used to create more friendly interaction between human and computer by the detection of affective computing. Other studies concerning facial expression detec-
Ubiquitous Computing Technologies in Education

In addition, for some labs with special purposes (e.g., precise instruments, biotechnology and medical science), it is necessary to detect the volume of particles in the air. Taking semiconductor as an example, these particles may cause instruments a short circuit and even disable the devices.

RESEARCH ISSUES AND TARGET SUBJECTS

In Taiwan, the Conference on Mobile and Ubiquitous Learning was initiated in 2006 (which is called Ubilearn’2006). Based on the issues discussed in Ubilearn’2006 and Ubilearn’2007, several potential issues have been taken into account:

1. Architectures and infrastructures for ubiquitous learning systems
2. Adaptive and adaptable learning environments using mobile and ubiquitous devices
3. Agent support for mobile and ubiquitous learning
4. Architectures and implementations of context-aware learning technology systems
5. Designs for wireless, mobile and ubiquitous technologies in education
6. Design of learner-supportive interfaces of ubiquitous learning applications
7. Evaluation and evaluation methodologies for ubiquitous learning
8. Entertainment computing for ubiquitous learning
9. Mobile and ubiquitous computing support for collaborative learning
10. New pedagogical theories for ubiquitous learning environments and society
11. Innovative and practical use of wireless, mobile and ubiquitous technologies for education, learning and training
12. Psychological or theoretical foundations for mobile and ubiquitous learning and training

Moreover, a special interest group on Mobile and Ubiquitous Learning Environment (MULE) was established in 2007. Based on the previous experiences from these researchers, some potential applications of context-aware ubiquitous learning have been given, including art courses (painting or drawing from life or nature), physical education courses (motor skill training), language courses (conversation training), natural science courses (plant and animal classification) and engineering courses (equipment operating and work flow training).

Nowadays, u-learning is gradually becoming an important learning style owing to the popularity of sensor and wireless network technologies. As long as we strive toward the aim and develop a feasible and effective learning model, the ideal learning environment will become reality. In this innovative learning environment, the system can more actively provide more adaptive assistance to the student based on their learning behaviors in the real world, which have revealed the coming of a new educational age.

ABSTRACT

Team-based learning is an active learning instructional strategy used in the traditional face-to-face classroom. Web-based computer-mediated communication (CMC) tools complement the face-to-face classroom and enable active learning between face-to-face class times. This article presents the results from pilot assessments of computer-supported team-based learning. The authors utilized pedagogical approaches grounded in collaborative learning techniques, such as team-based learning, and extended these techniques to a Web-based environment through the use of computer-mediated communications tools (discussion Web-boards). This approach was examined through field studies in the course of two semesters at a US public technological university. The findings indicate that the perceptions of team learning experience such as perceived motivation, enjoyment, and learning in such a Web-based CMC environment are higher than in traditional face-to-face courses.
In addition, our results show that perceived team members’ contributions impact individual learning experiences. Overall, Web-based CMC tools are found to effectively facilitate team interactions and achieve higher-level learning.

INTRODUCTION

Instructors of both traditional face-to-face and online classrooms seek active learning techniques that engage the learners. The increased use of Web-based computer-mediated communications (CMC) as support tools that supplement the face-to-face classroom (“blended learning”) and enable active learning between face-to-face class times fit this quest. CMC is regarded as an efficient computer support tool to facilitate student participation (Phillips & Santoro, 1989). Prior research (Wu & Hiltz, 2004) reports that adding asynchronous online discussions through CMC platforms enhances students’ learning quality in a face-to-face class setting. Although various Web-based computer-mediated communications learning strategies have been applied in the field (e.g., online collaborative learning), limited research focuses on computer-supported team-based learning in a face-to-face classroom. Team-based learning (TBL) is an instructional strategy that promotes active learning in small groups that form a team over time (Michaelsen, Fink, & Knight, 2002).

Our goal is to assess the impact of team-based learning when introduced in a face-to-face classroom that utilizes Web-based CMC as a supplemental learning tool between classes, thus increasing team interaction across the semester. A Web-based computer-mediated communications tool called WebBoard™ was utilized in our computer-supported team-based learning research to facilitate team learning activities and communication. This paper describes results from this experience. The paper begins with a literature review building on constructivist learning, collaborative learning, small group learning, and Bloom’s taxonomy theories. It then provides examples of the Web-based interface and pedagogical implementation, introducing a model for assessing computer-supported team-based learning. Research questions, hypotheses, and data analysis results are presented. Finally, the limitations of the study and future research efforts are discussed.

We believe that our contribution is two-fold. First, we describe an approach for transferring a grounded pedagogical approach to a Web-based environment by supplementing the experiences from a face-to-face classroom. Second, we document preliminary assessment results that support the feasibility and effectiveness of the proposed approach. This discussion should be of interest to educators and researchers in expanding the use of current Web-based learning management systems with a structured modular approach through the integrated use of discussion forums to achieve higher-order team learning outcomes.

THEORETICAL BACKGROUND

Constructivist Learning Theory

Leidner and Jarvenpaa (1995) classify learning models and discuss their relevance and impact in information systems educational approaches. The broadest categories of this classification are objectivism and constructivism. Objectivism posits that learning occurs in response to an external stimulus. Learners respond to the stimulus by modifying their behaviors. This model assumes that abstract representations of reality and knowledge exist independently from the learners. Teaching consists of transferring knowledge from the expert to the learner. Opposite to objectivism, constructivism posits that learning is not a process of knowledge assimilation, but an active process of constructing individual mental models, in which knowledge is created.
in the mind of the learner. In this model each individual controls the pace and depth of his/her instruction. The instructor is only a moderator in the process of hypothesizing, questioning, and discovering the conceptual relationships between and among various objects.

Team-based learning uses a constructivist approach, which converts the learner from a passive to active learner. This differs from the traditional teacher-learner or objectivist approach. Students in this constructivist learning environment play a more active role as learners, since they need to be well-prepared in order to effectively engage in various class activities, such as to facilitate class discussions, to be able to take challenges from their peer learners and instructors, and so forth. Therefore, the constructivist approach is aimed to facilitate students’ critical thinking to achieve higher-level learning.

This research explores how the constructivist approach can be utilized to promote team-based learning in a Web-based CMC tool environment. Building on experiences from face-to-face team-based learning teaching (Michaelsen et al., 2002), computer-mediated communication tools are adopted to expand the collaboration opportunities via online team activities between weekly classes. Suitable technologies used in this context are listservs (email distribution lists with a Web-based interface) and asynchronous learning network systems (e.g., WebBoard, WebCT, BlackBoard, or similar technologies provide both a synchronous and an asynchronous discussion environment).

Collaborative Learning

Built upon the constructivist theory, collaboration suggests that learning quality enhances through shared understanding of more than one learner. Due to social interactions and intensive sharing among collaborative learners, collaborative learning results in higher learning compared with individual-oriented learning (Leidner & Fuller 1996). Schlechter (1990) found that learners perceived higher-order learning by generating more creative ideas and more diverse reasoning in collaborative small study groups.

Our computer-supported team-based learning approach takes advantage of collaborative learning strategies, with more structured teams, which exist for a relatively longer period (in our case, in the entire semester) and with a stable team setup (i.e. same team members), in comparison to other approaches where most teams are built much more casually and temporally.

Small Group Learning and Team-Based Learning

The importance of small groups learning and knowledge creation has been increasing in both education and industry. Team-based learning focuses on fixed small groups, which are established for semester long collaboration instead of temporary purposes. Small groups promote each other’s learning and success by holding each member personally responsible for the fair share of the work (Johnson, Johnson, & Smith, 1991). This turns the learning experience into a process, which improves the quantity and quality of the learning by leveraging long-term caring and peer relationships (Johnson & Johnson, 1999). Team-based learning simulates the similar peer collaboration experience in a real world. Therefore, the team-based learning experience may represent a useful training for students’ long-term career success.

The main emphasis of team-based learning is the organization around modules (work units) across the semester, consisting of 5-7 three phase sequences (Michaelsen et al., 2002). Each sequence includes preparation, application and assessment before moving to the next unit. Teams should be five to seven members in size. They evolve through four essential procedures: team formation, student accountability, team activi-
ties, and high quality feedback (Michaelsen et al., 2002). Our research uses these four procedures as the basis for the introduction of computer-supported techniques.

**Phased Cognitive Development Perspectives**

The approaches presented in this study are based on pedagogical theories that recognize the role of incrementally promoting higher-level learning through progressive team activities. Bloom, Englehart, Furst, Hill, and Krathwohl (1956) organize learning based on different phased cognitive development stages of information elaboration, starting from knowledge and comprehension to the higher levels of evaluation and synthesis. Bloom et al.'s taxonomy of learning goals and objectives summarizes the way student learning progresses through various stages of increasing complexity.

Bloom's taxonomy is particularly useful to identify a progression of learning, which is applied to the computer-supported team-based learning process. In the lower levels of the taxonomy (knowledge, comprehension, application) students absorb knowledge that is presented to them, comprehend it, and learn how to rank order and make inferences in other contexts. In the higher levels of the taxonomy (analysis, synthesis, and evaluation), students learn to generalize and transfer what they learned to their personal interpretation of reality. They learn to analyze and classify the models they have created. While the synthesis and evaluation levels of the taxonomy can also be accomplished by individual learners, collaborative methods and working in teams provide opportunities to excel in integrating associations and making judgments through communication with peers. In this collaboration mode, students can reach the higher levels of the Bloom’s taxonomy.

**WEB-SUPPORTED TEAM-BASED LEARNING EXPERIENCES**

Computer-supported team-based learning conducted via a Web-based CMC tool was introduced in a 15-week semester, face-to-face graduate course at a US public technological university during the Fall 2004 and Spring 2005 semesters. The first class period introduced the computer-supported team-based learning instructional strategy. The class was divided into teams of five to six students, including at least one woman and at least one student with a wireless laptop. The course materials were divided into six modules, with no midterm or final. Two out-of-class individual article reviews were assigned to the students.

The graduate courses in which we conducted this research was re-modeled, according to Michaelsen et al. (2002). Several steps of the TBL instructional strategy were modified to incorporate the use of a Web-based CMC tool called WebBoard (http://www.webboard.com/) intended for active learning activities between weekly face-to-face classes. A phased approach to active learning assignments via WebBoard was implemented across the two semesters of the study (Gomez & Bieber, 2005). Particular emphasis was placed on additional individual preparedness pre-module activities and team post-module activities through the use of WebBoard; a learning management system that provides a flexible interface for classroom discussion management. Figure 1 shows a typical discussion interface in WebBoard. The left frame provides an organized list of class activities. The right frame shows the details of each individual posting. For example, in the first semester, students posted preparation materials on WebBoard (http://www.webboard.com/) and wrote their post-team activity results on a poster in class. In the second semester, teams posted results of their work on WebBoard instead of handwriting on posters.
The team-based learning activities are based on an iterative process that divides the course into five to seven modules (themes/work units), as Michaelsen et al. (2002) suggests. A module is defined as a work unit or theme. Each module introduces increasingly complex materials that students must master before a subsequent module is introduced. A module can span more than one week’s worth of course materials, but a single activity should not span more than one class. Web-based CMC tools enable spanning beyond the single class time (when materials are hard to understand) by enabling the team to work between scheduled class times. Each of the modules must then be subdivided into individual preparation (reading materials and discussions), readiness assessment tests (individual and team), multiple team activities, and post module assessment.

The instructor acts as a facilitator across each module. He/she is responsible for hypothesizing, questioning, and discovering the conceptual relationships between and among various concepts (Passerini & Granger, 2000), and spending more time on preparation of the readiness assessment tests (RAT) and team activities, rather than course lecturing. The instructor plays a pivotal role in the TBL process, both for the course organization and ongoing module activities. In lieu of class lecturing, the instructor actively participates in the classroom as a task coordinator and time manager while observing each team’s discussions to readily intervene either as a subject matter expert or to clarify points of the task. Timely grading and feedback are essential aspects of the TBL process and should not be overlooked when planning. The instructor also provides feedback when team activities come together for sharing across the classroom.

Our emphasis with the introduction of Web-based tools to the TBL process is on the individual

![Figure 1. WebBoard interface](image-url)
Utilizing Web Tools for Computer-Mediated Communication to Enhance Team-Based Learning

**Table 1. Team-based learning instructional strategy**

<table>
<thead>
<tr>
<th>TBL Elements</th>
<th>Recommended (Michaelsen, 2002)</th>
<th>Description and Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Course Structure</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructor Lecturing</td>
<td>Discouraged</td>
<td>The instructor acted as a facilitator and observed discussions. No lecturing was provided.</td>
</tr>
<tr>
<td>Individual Preparedness</td>
<td>Assign all module reading</td>
<td>All reading materials were assigned, reading summaries, and test notes were posted on WebBoard before the RATs.</td>
</tr>
<tr>
<td></td>
<td>materials at start of module.</td>
<td></td>
</tr>
<tr>
<td>Readiness Assessment</td>
<td>At the start of each module:</td>
<td>Six modules, each team can ask one question about module materials, individual test (tRAT) followed by team test (tRAT).</td>
</tr>
<tr>
<td>Test (RAT)</td>
<td>5-7 modules per semester.</td>
<td></td>
</tr>
<tr>
<td>Team Activities</td>
<td>Multiple activities of increasing</td>
<td>Multiple team activities were given with increasing complexity per module.</td>
</tr>
<tr>
<td></td>
<td>complexity per module.</td>
<td></td>
</tr>
<tr>
<td>Grading</td>
<td>Tests, assignments and certain</td>
<td>Class determined grade weightings of all exams/tests, assignments and activities.</td>
</tr>
<tr>
<td></td>
<td>in-class team activities should</td>
<td></td>
</tr>
<tr>
<td></td>
<td>be graded.</td>
<td></td>
</tr>
<tr>
<td>Course Survey</td>
<td>Not mentioned.</td>
<td>End of semester.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Team Formation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team Size</td>
<td>5-7</td>
<td>5-6 students for each team.</td>
</tr>
<tr>
<td>Team Arrangement</td>
<td>Same groups</td>
<td>Same groups working together all semester.</td>
</tr>
<tr>
<td>Team Roles</td>
<td>No assigned roles.</td>
<td>Roles not assigned.</td>
</tr>
<tr>
<td>Team Composition</td>
<td>Not specified.</td>
<td>Balanced distribution among gender.</td>
</tr>
<tr>
<td>Team Diversity</td>
<td>Recommended that diverse</td>
<td>Students were assigned to teams based on their computer experience.</td>
</tr>
<tr>
<td></td>
<td>groups are formed randomly.</td>
<td></td>
</tr>
<tr>
<td>Team Building</td>
<td>Achieved through initial tRAT,</td>
<td>As recommended in addition to a phased approach to within team reading summaries and sharing of notes for tests.</td>
</tr>
<tr>
<td></td>
<td>team exercise to determine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>grade weights and choose a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>team name.</td>
<td></td>
</tr>
</tbody>
</table>

preparedness of the student before taking the individual readiness assessment test for a given module and subsequently taking the team readiness assessment test. To accompany the reading assignments, students were asked to post a short reading summary and then reply to another student’s reading summary. In the Fall 2004 semester, the reading summaries were discussed online across the classroom. In the Spring 2005 semester, the reading summaries were discussed within each team. Our observations across both semesters indicate that discussions within the team demonstrated two benefits: extended discussions with the team and increased discussions prior to the team readiness assessment tests, reinforcing the preparation at the start of a new module. This suggests that discussions between team members are more effective and rewarding than discussions.
across the entire classroom. In reviewing the discussion board postings, the length and depth of the discussions were richer and more engaging when individuals were discussing within their team. This provides the additional benefit of increased team cohesion; a quantitative measure to be included in future research.

The student transition from their individual preparedness activities to the readiness assessment tests. Individual preparedness activities include all reading materials for the module, reading summaries, and individual exam notes for sharing with teammates, requiring students to organize their preparation time. It is suggested the questions for the readiness assessment tests should span the modules reading materials and be geared toward concepts; not detailed questions typically categorized as “by rote”. The questions should be broad enough to cover all of the module’s reading materials. Michaelsen et al. (2002) suggests approximately twenty multiple choice questions at the start of each new module for the readiness assessment test (RAT). Our research used a point system for the team readiness assessment test where teams were given a scratch card and were allowed to select more than one answer. Points were awarded in decreasing value (e.g. five points first try, three points second try, one point third try). Teams are permitted to appeal a question, with five to ten supporting sentences to ensure a concise and well thought out rebuttal. Students build cases with convincing evidence to persuade the teacher to award credit for an answer missed by the group on the test (Michaelsen et al., 2002).

Multiple team activities per module are needed with increased levels of complexity and are not assigned until after the team readiness assessment test. Typically team activities span two or three weeks for a single module, but a single activity does not carry over from one week to the next.

**Figure 2. Computer-supported team-based learning events**
The ideal activity is one that is a case study or has a problem-solving aspect to the activity. After each team activity, deliverables were reviewed and critiqued by the other teams, possibly for extra credit. The review and critique occurred in the FtF classroom for the first semester and using WebBoard for the subsequent semester. The premise was to deepen the learning and reinforce the objectives of the activity, while providing the instructor an opportunity to comment. Upon completion of a team activity, students should review the other teams’ deliverable and provide feedback to generate discussion.

Figure 2 summarizes the team-based learning iterative process and where computer-supported activities were introduced in this study.

ASSESSING COMPUTER-SUPPORTED TEAM-BASED LEARNING

The constructivist approach, cooperative, and collaborative theories and models, provide the background for the elaboration of the team-based learning assessment framework reviewed in this study (see Figure 3). Earlier research on Web-supported participatory examinations (Wu, Bieber, Hiltz, & Han, 2004) and collaborative examinations (Shen et al., 2004), informs the proposed team-based learning constructs and expected relationships. In order to succeed in this TBL process, students should prepare first, so that they can contribute their insights to the whole team (*individual preparedness* construct). We assume that their individual preparedness might bring more confidence for them on the value their team member’s contributions to problem solving, and they can also integrate and gain more insights from their team members. During this TBL process, students can learn more by communicating efficiently, and by building trust among their team members. Overall, we expect that the whole TBL process will provide the students an enjoyable learning experience, and will motivate them think at the higher levels of Bloom’s learning taxonomy, enhancing the quality of the learning experience. Figure 3 lists the key constructs and hypotheses tested in this study.

*Figure 3. Team-based learning (TBL) framework*
Definitions: Perceived Individual Preparedness is a self-reported student assessment of deep versus superficial study of the materials (Gomez, Wu, Passerini, & Bieber, 2006; Michaelsen et al., 2002); Perceived Team Member Value or Contributions is the positive perception of the value of teamwork (Isabella & Waddock, 1994); Perceived Trust/Communication Skills refers to the openness of communication within the teams (Coppola, Hiltz, & Rotter, 2004); Perceived Motivation refers to the intrinsic and extrinsic factors that influence students’ interest and attitudes towards the learning experience (Malhotra & Galletta, 2003); Perceived Enjoyment is the extent to which the learning activity (the team-based learning experience) is pleasant and satisfactory to the learners (Davis, Begozi, & Warshaw, 1992); Perceived Learning focuses on perception of the course quality; usefulness and extend of individual learning experiences as reported by the learners (Wu et al., 2004).

In exploring the relationships and impacts of computer-supported team-based learning, we focused on a number of factors that could impact students’ perceptions of learning. We adopted and modified a few validated constructs—“perceived learning,” “perceived motivation” and “perceived enjoyment”—from asynchronous online discussion research (Wu & Hiltz, 2004). In addition, we created two new constructs called “perceived individual preparedness” and “perceived team members’ value/contributions.” More specifically, we expect that the collaboration experiences and team-based activities completed by the teams throughout the courses will have a positive impact on perceived learning. Individual preparedness, measured as a self-reported student assessment of their deep versus superficial study of the materials, will positively impact the perceived value and contribution of the team-learning experience. This assumption follows Michaelsen et al.’s (2002) findings that individual contributions to team-output will promote team development and reduce social loafing and, therefore, will have a higher impact on the overall team-based learning experience.

Prior asynchronous learning networks research (Wu & Hiltz, 2004; Wu et al., 2004) shows a positive correlation between perceived motivation, enjoyment, and learning, when students participate in designing, answering, and grading exam questions on an asynchronous learning networks tool. These findings occur in the context of students’ learning experience from online discussions through computer-mediated communications platforms for blended classes, which mixed traditional classroom lecturing with asynchronous online discussions.

Individual preparedness and perceived member contribution to computer-supported team-based learning are expected to have a positive impact on learning, as well as on motivation and enjoyment of the team-based learning experience. In particular, motivation and enjoyment will act as intervening variables (see Figure 3). We also expect motivation to positively impact the overall enjoyment of the team-based learning exercises. Following Thorndike’s “Law of Intensity” and “Law of Readiness” (Thorndike, 1932), we expect that students who are engaged in the learning process through the multiple collaboration experiences embedded in the course organization, will be more motivated to learn, and will enjoy the learning experience better.

Lastly, we expect that the independent variables of individual preparedness, perceived team member value and contribution to computer-supported team-based learning will impact the level of trust and openness of communication within the teams, indirectly influencing communication and enjoyment. Establishing trust early (swift trust) in on-line communities has been found to have a positive impact (Coppola et al., 2004) on the learning experience. While we do not report results on trust measures (further analysis and
reiteration are needed as trust measures were collected only in the Spring 2005 reiteration of the field study, we include trust and communication as an important variable to consider in a computer-supported team-based learning model, which will be influenced by the individual preparedness and team contributions. As discussed in the future research section of this paper, additional observations in the Summer and Fall semester 2005 will supplement our preliminary observations.

Research Questions and Hypotheses

Based on the above discussion, our research is set out to investigate the following general questions:

- Does students’ perception of team contributions impact their learning from the computer-supported team-based learning process?
- Does individual preparedness affect perceptions of computer-supported team-based learning experiences?

A number of research hypotheses are therefore derived from the proposed research framework (see Figure 3).

**Hypothesis 1a&b:** Higher individual preparedness will increase students’ perceived motivation and enjoyment from computer-supported team-based learning process.

**Hypothesis 1c:** Higher individual preparedness will increase the perception of team members’ value and contribution to the computer-supported team-based learning process.

**Hypothesis 2a&b:** Higher perceived team members’ contributions to computer-supported team-based learning will increase perceived motivation and enjoyment from this process.

**Hypothesis 2c:** Higher perceived team members’ contributions to computer-supported team-based learning will enhance perceived learning in this process.

**Hypothesis 3a&b:** Higher perceived motivation will lead to higher enjoyment and learning from the computer-supported team-based learning.

**Hypothesis 4:** Higher perceived enjoyment from the computer-supported team-based learning will lead to higher learning.

Methods and Sampling

To test the hypotheses, survey questionnaires were used in the same masters-level information systems course called “Information Systems Principles” during Fall 2004 and Spring 2005 semesters. A total of 73 students volunteered to participate in our computer-supported team-based learning study. Among 73 respondents, 61 students disclosed their demographic information. Over 50% of them are full-time students, about 33% are part-time, and the rest are non-matriculated students, who are studying without officially being admitted to degree programs. About 60% are males, 36% are females, and three people did not provide their gender information. The majority of the respondents are between 21 to 30 years old, about 16% are between 31 to 40 years old, and the rest are between 41 to 50 years old.

Data Analysis and Discussion

In this study, the majority of data was collected from the surveys, which covered all variables proposed in the computer-supported team-based learning research framework (Figure 3). The primary data analysis methods to analyze our survey are descriptive analysis, factor analysis, and correlation analysis. First, descriptive data analysis was conducted to evaluate the validity of the constructs and the reliability of the scale used in the study. All the constructs returned a
Cronbach’s Alpha higher than 0.70. The detailed question items for each construct are listed in Appendix section.

The factor analysis was then performed to confirm the loading of the questions to the respective factors. In general, the question items show high reliability of the constructs they represent. The exploratory factor analysis technique was used in this study to verify whether the grouping of the survey questions maps to the model constructs.

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial Eigenvalues</th>
<th>Extraction Sums of Squared Loadings</th>
<th>Rotation Sums of Squared Loadings</th>
<th>% of Variance</th>
<th>Cumulative % Variance</th>
<th>Total</th>
<th>Cumulative Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.612</td>
<td>37.442</td>
<td>37.442</td>
<td>8.612</td>
<td>37.442</td>
<td>37.442</td>
<td>8.612</td>
</tr>
<tr>
<td>3</td>
<td>1.721</td>
<td>7.484</td>
<td>55.716</td>
<td>1.721</td>
<td>7.484</td>
<td>55.716</td>
<td>36.577</td>
</tr>
<tr>
<td>4</td>
<td>1.518</td>
<td>6.602</td>
<td>62.319</td>
<td>1.518</td>
<td>6.602</td>
<td>62.319</td>
<td>56.545</td>
</tr>
<tr>
<td>5</td>
<td>1.122</td>
<td>4.879</td>
<td>67.197</td>
<td>1.122</td>
<td>4.879</td>
<td>67.197</td>
<td>64.807</td>
</tr>
<tr>
<td>6</td>
<td>1.034</td>
<td>4.497</td>
<td>71.694</td>
<td>1.034</td>
<td>4.497</td>
<td>71.694</td>
<td>71.694</td>
</tr>
<tr>
<td>7</td>
<td>1.034</td>
<td>4.497</td>
<td>71.694</td>
<td>1.034</td>
<td>4.497</td>
<td>71.694</td>
<td>71.694</td>
</tr>
<tr>
<td>8</td>
<td>8.57E-02</td>
<td>363</td>
<td>100.000</td>
<td>8.57E-02</td>
<td>363</td>
<td>100.000</td>
<td>100.000</td>
</tr>
</tbody>
</table>

Table 3. Components extraction matrix
An exploratory factor analysis (EFA) was run to identify the underlying model, the number of factors in the model, and the variables (by type of questions) associated with each factor. Within the factor analysis technique, a principal components analysis (PCA) was used to highlight the number of factors to be extracted from the model. The confirmatory factor analysis (CFA), in its rotated version (VARIMAX rotation) helps to identify the underlying model. The Principal components analysis yielded six factors with eigenvalues >1. We focused on the first five components corresponding to the research model constructs (Table 3) because of the limitation of one of the research constructs (individual preparedness) that yielded ambiguous results and because the additional variance explained by the additional factor did not substantially alter the model (as it is the lowest incremental cumulative variance). In addition, we did not test the trust/communication construct, with a possible impact on the overall model. Limitations of this approach as introduced at the end of this section.

**Extraction Method:**
**Principal Component Analysis**

Table 4 shows the bivariate correlations among variables and provides intriguing results. In particular, it shows that perceived enjoyment and motivation are highly correlated with the dependent variable of perceived learning (R>0.63 for both constructs). It also shows that perceived team member’s contribution positively correlates to learning, enjoyment and the motivation constructs. However, the bivariate correlation analysis shows that individual preparedness is not significantly related to the other variables, and the relationship, if any, is negative.

Based on the results of the above bivariate correlations, we identify that hypothesis 1 (a, b, and c) are not supported. The other hypotheses are supported at the p=0.01, with the correlation between team contributions and motivation significant at the p=0.05 level (see Figure 4).

From the above data analysis results, it is suggested that how individuals value their team
Utilizing Web Tools for Computer-Mediated Communication to Enhance Team-Based Learning

members’ contribution has significant correlations with their perception of enjoyment and learning quality from computer-supported TBL (Gomez et al., 2006). These results address the first research question on how team interactions positively impact the whole computer-supported TBL learning experiences. In addition, individual opinions on team members’ contributions also have a positive impact on their perceived motivation, although the Pearson’s R value is not high (R = 0.28). It might be caused by other potential factors from the computer-supported TBL experience, which could decrease students’ motivation. For instance, if the team leader is more dominant, his control might impact other team members’ motivation.

Surprisingly, individual preparedness does not impact perception of the team-based learning experience. The correlation values among individual preparedness and other variables are not significant. The results also show that their correlation values are negative. There might be a few reasons. First, as indicated earlier, the ambiguous two question items for the “individual preparedness” construct (i.e. ambiguous statement, negative factor loading) have a negative impact on the research framework. The individual preparedness tasks associated with studying differences may not be self-evident to the student increasing the ambiguity of the questions. Moreover, our observations indicate the importance of individual preparedness on the entire iterative nature of the module whereas our focus of the current individual preparedness construct was driven from the importance of the readiness assessment test process. The bivariate correlation analysis does not show any significant relationships between “individual preparedness” and other constructs in the framework. This is evidently one of the limitations for this research. For future research, we plan to refine the “individual preparedness” construct with question items targeted at the individual preparedness process more than studying specifically for the tests. Second, there might be an interaction effect of the experimental conditions: the computer-supported TBL process design itself might also impact the results. The team assessment tool (tRAT) is the same test as the individual readiness assurance test (iRAT). Although the overall team scores are better than individual test scores, the test repetition may explain the decrease in the students’ motivation and enjoyment. Also they could not perceive more value from their
Utilizing Web Tools for Computer-Mediated Communication to Enhance Team-Based Learning

respective team members. We also speculate that the test questions may not have lent themselves well to interesting team discussions, leaving the tRATs uninteresting. Alternatively, this simply may indicate that many students found the TBL process valuable even when they did not prepare in the manner the instructor expected. These results show that our second research question on the role of individual preparedness needs further investigation and analysis to span the iterative cycle of each module (preparation, test, activity) and not simply focus on the readiness assessment test aspect.

There are a number of limitations to this study. First and foremost, we found that some of the constructs could be better specified. Our factor analysis displayed the possible existence of another significant factor which should be further analyzed in future research. This additional factor was associated with an ambiguous question in the individual preparedness construct. Looking at the articulation of the questions in that construct, we identified an instrumentation bias (one item in the construct was weak), which may explain our concerns with the specific construct. In addition, because some of the constructs and the extension of this grounded team-based learning approach to a Web-environment is novel, we should supplement our conclusions with an in-depth analysis of qualitative factors (observations, open-ended questions, and content analysis of discussion boards) that may help better understanding the learning outcomes more objectively in a way that complements the perceived learning measures. Adjusting the activities layout in the Web-based environment with our phased approach and the nature of the individual preparedness activities indicate the complexities of this construct. Our qualitative analysis indicates that spanning beyond the individual preparedness question items related to the readiness assessment test to also include question items that measure perceived preparation of the materials related to team activities.

SUMMARY AND FUTURE RESEARCH

Computer-supported team-based learning provides a powerful instructional experience and reduces some of the disadvantages many instructors and students have found with traditional small-group work. Blending the benefits of the face-to-face classroom with computer-mediated communication extends the learning process between the weekly face-to-face sessions keeping the team learning progress and group dynamics growing.

Our research places emphasis on key variables that affect learning in computer-supported team-based learning. Computer-supported team-based learning is still a relatively new pedagogical approach, and to the best of our knowledge, this is the first study blending computer-mediated communications with the iterative team-based learning modular approach proposed by Michaelsen et al. (2002). The use of Web-based CMC learning techniques places emphasis on individual and team learning outcomes. The surveys indicate a high-perception of learning, motivation, and enjoyment. These findings deemed computer-supported team-based learning an approach for further investigation both in the face-to-face classroom and for online learning.

The emphasis of future research will be on team-assessments and group cohesion in a purely Web-based learning environment. The findings around the team activities will allow for additional adjustments in the team-based learning process before it is introduced in a completely online learning mode. Blending the face-to-face class with computer-mediated communications provides a means to gauge the asynchronous learning network process. Future studies will extend the analysis of the computer-supported team-based learning model and research framework using the structural equations model (SEM), trust, com-
Utilizing Web Tools for Computer-Mediated Communication to Enhance Team-Based Learning

munication, and team leadership factors. Further review of individual and team preparedness is also needed. The progressive nature of the readiness exam process and team activities should ensure individual preparation. Because of the novelty of the preparedness and team contributions constructs, we will also implement content analysis of team activities posted on WebBoard to support the evaluation of individual preparedness for each module. Adding qualitative data and observations will enhance our understanding of the constructs. Actual grades and peer evaluation results will also support the measurement of task completion levels.

Team-based learning presents a promising technique employing small teams that actively engage students in learning. We look forward to the day when instructors can effectively use computer-supported team-based learning as a standard approach in both face-to-face and online classrooms.

ACKNOWLEDGMENTS

We gratefully acknowledge partial funding support for this research by the United Parcel Service Foundation, the New Jersey Center for Pervasive Information Technology, the New Jersey Commission on Science and Technology, and the National Science Foundation under grants IIS-0135531, DUE-0226075 and DUE-0434581, and the Institute for Museum and Library Services under grant LG-02-04-0002-04. An earlier version of this paper was presented at the IRMA 2006 International Conference in Washington DC, May 2006.

REFERENCES


*This work was previously published in the International Journal of Web-Based Learning and Teaching Technologies, edited by L. Esnault, Volume 2, Issue 2, pp. 21-37, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 3.24
How Technology Can Support Culture and Learning

David Luigi Fuschi
GIUNTI Labs S.r.l., Italy

Bee Ong
University of Leeds, UK

David Crombie
DEDICON S.r.l., The Netherlands

ABSTRACT

From the authors’ perspective, technology is both a tool and a developing factor that can foster culture and learning development. This chapter focuses on the interrelations that interleave education, technology, content accessibility, and intercultural issues. With an introduction and related background, language learning is used as an example further to explain these issues. This chapter then discusses authoring and content development for e-learning applications (including authoring tools, virtual communities, and forums), and examines technology and accessibility issues in this context. The current state of e-learning is introduced along with a description of different tools and approaches. The chapter concludes with an overview of e-learning and the marketplace.

INTRODUCTION

Major developments in culture and civilisation are based on a combination of factors including personal abilities, societal, and economical environment, and so on. According to Maslow (Abraham, 1943), subsistence needs have to be met before other and “higher” aims are reached for (Figure 1). Nevertheless, the search for a solution to a problem (even a very practical one) may provide the core of our evolution in technology, science, culture, art, or education.

From the authors’ perspective, the issue regarding the relation interleaving technology, culture, and education is twofold. On one side, technology has often been a major cause of disruptive advances in the evolution of science, culture, art, and education. On the other side, scientific
How Technology Can Support Culture and Learning

For example, the present development of western culture has been reached thanks, in large part, to the technical innovation introduced by Gutenberg, which made possible the diffusion of culture through print, making learning affordable for everyone and therefore, also accessible. Yet were it not for Braille, all the culture and knowledge available in printed form would be totally inaccessible to visually impaired people unless they could afford to have a personal reader. This technical support is therefore a clear example of how technology can support culture and learning in the wider sense.

It is worth taking into account historical notes related to the impact of technology on the learning process and therefore, also on the evolution of society.

**SOME HISTORY NOTES**

The art of teaching and learning has a long history. Initially it was a matter of subsistence and therefore, the elders in a community would act as tutors, trainers, and mentors for the younger members. This process then evolved from mere subsistence to something that became much valued by the learner and rewarding for the teacher.

Due to the very nature of this process (based on knowledge acquired through personal practical experience), teachers were numerically very...
limited in respect to potential learners; moreover, not everyone could benefit from mentoring, as most of the individuals would have to work hard to support the group’s living needs. Therefore, only the most brilliant, or the more influential members, would be granted such a precious gift as “formal learning” and this had a strong social impact.

For a long period of time, education and culture has been accessible only to upper classes; religion played a relevant role in this, as knowledge was mainly held by “priests” or “monks,” who would also act as tutors and mentors for nobles, ruling classes, and their earls or illegitimate sons.

This phenomenon was so strong that during the “Middle Ages” the term “clerices” was used generically to indicate the people with an education. This is also the period of the birth of the first universities like the “Alma mater” in Bologna (1088), the “Sorbonne” in Paris, “Montpellier,” the “Scuola Medica Salernitana,” “Heidelberg,” “Salamanca,” and many other ancient and prestigious universities in Europe. The “clerices vagantes,” in their roaming through Europe, spread culture, teaching in the various universities they visited, rivalling each other to achieve eternal fame thanks to the universal language of the time (Latin). A famous example is Abelard (the well-known character of Abelard and Eloise). Most people consider this as a tale while it actually was a real event in European history.

The main academic subjects at the time would have been literature, philosophy, religion, and law, while science was more limited in scope than now. In this sense, it is interesting to find out how some of the presently most reputed universities started, for example Oxford and Cambridge. Their origins date back to 1167 when Henry II ordered all English students on the Continent to return to England. Many of them decided to settle in Oxford in an attempt to create a university as they had seen in Europe. Unfortunately, disputes between the students and residents led to riots and a number of students fled to Cambridge, where they established a new university. In turn, the loss of students hurt Oxford’s local economy and in 1214, traders and merchants invited them back again. The competition between Oxford and Cambridge colleges probably dates back to those times when William of Durham founded University College in 1249. Balliol was established in 1263 and Merton, the first residential college, followed a year later. Merton set the collegiate pattern, which became standard in both Oxford and Cambridge. These colleges were self-governing institutions where teachers and students lived together. Fourteen other colleges were founded by the end of the 16th century (http://www.spartacus.schoolnet.co.uk/EDoxford.htm).

In another part of the world, the “clerices” role was covered by other scholars; it is remarkable that most of them would still be monks, hermits, or anchorites like Buddhist and Confucian scholars in Asia, or Imams and Ayatollahs in the Muslim world. Moreover, to most people, the universities previously mentioned seem to have had an extremely early start, but if we look at the records of Nalanda, we find out that this is actually mistaken.

Nalanda was a Buddhist monastic centre, often referred to as a university, in the North Indian state of Bihar. Though it is traditionally placed in the time of the Buddha (6th—5th century B.C.), archaeological excavations date its current foundations to the 5th century A.C.; it housed a population of several thousand teachers and students. The topics covered spanned logic, grammar, astronomy, and Ayurvedic Buddhist medicine; the Chinese pilgrims Xuanzang and Yijing provide vivid accounts of Nalanda as early as the late 7th century. Nalanda continued to flourish through the 12th century and became a centre of Buddhist religious sculpture, but was probably sacked during Muslim raids around 1200, and has never recovered since (http://www.ayurveda-tcm.com/Buddhist_ayurveda_monastic_college_Nalanda.htm).
In China, the First Emperor (259 B.C.) established a process to formalize a set of hierarchies and exams related to the acquisition of public roles (becoming a Mandarin or an emperor’s officer). In similar fashion measures, rules, laws, and even the way of writing were standardized and formalized so that the same characters are still in use. Since then, in China, culture has always been highly valued; there is even a common saying related to this that translates as “a word is worth a 1,000 pieces of gold.”

The origin of this common Chinese saying can be found in the famous book Shiji (or Historical Records) by the historian Sima Qian (around 145-87 B.C.). He describes an episode from the time when Lü Buwei (the natural father of the first emperor) was prime minister. Lü Buwei was hosting and protecting over 3,000 scholars, so he compiled the best of their writings in a book, claiming that it was the encyclopedia of his time covering science, literature, and all the knowledge that was under heaven. Actually, the prime minister did more; he had a collection of texts engraved onto rock and an edict proclaimed stating that: “Whoever would have been able to improve all that is collected and written by simply adding or removing even a single word would have been granted a 1,000 pieces of gold.” For fear of offending the prime minister, a very powerful man, no one dared to attempt this, but since then it was clear why even a single word would be worth a fortune (http://www-chaos.umd.edu/history/ancient2.html, http://www-chaos.umd.edu/history/imperial.html#first, http://www-chaos.umd.edu/history/toc.html)(Mah, 2003).

Coming back to Europe, culture and education started to spread only with the rise of the middle classes and the development of the publishing industry, even though remaining the prerogative of a minority. Only in the last century, well in line with Maslow’s theory (Abraham, 1943), did education become a patrimony “for everyone.”

Figure 2. Evolution in educational models and results in relation to technology innovation
How Technology Can Support Culture and Learning

This is mainly due to the social and economic evolution that occurred after the industrial revolution, bringing better living conditions.

In the process of providing access to education and culture, technology had a crucial role. For example, the evolution in educational models and achievable results (in terms of educational objectives) in relation to technology innovation, along with major driving forces, is presented in Figure 2. In the diagram, the role played by economical forces as a push factor (companies demanding new learning services in order to have a more highly skilled work force) is shown. At the same time there were also socio-political forces pushing for more flexible and decentralised approaches, both in work and in education, to keep on track with society’s evolution.

In all this, technology was playing a twofold role. On one side, it was a driving force providing solutions to issues and needs arising from market and production environments. On the other side, many spin-offs of this evolution were also fed back into everyday life. Education and training could also benefit from technologies by providing new tools and methods to convey content and knowledge; thanks to the introduction of factors like simulation and interactivity at first, and then collaboration and personalisation, dramatic developments were achieved. At present, the new evolutionary step will likely be lead by the effective combination of these factors and the introduction of new technologies.

To better understand this, we will look at the origin of open and distance learning (ODL). This began in areas where the educational system was facing huge challenges due to distances and the spread of population over territory, combined with a growing demand for access to education, as in the USA, Australia, and Canada. To meet the requests (coming from all levels of society) for a higher level of education/training, a form of “distance learning” needed to be developed. This new learning process was initially supported by regular mail, then by radio and television. Then tools initially developed to support totally different scopes, like the Internet, started to be used for scientific and educational purposes. The initial spread of data among researchers fostered the exchange of knowledge; this brought on the rapid evolution of the so called “wired-world” that nowadays plays a crucial role in education, yet all previously mentioned media are still in use depending on needs and available budgets. As a matter of fact, available solutions cover almost all forms of training and teaching, from traditional class learning to e-learning.

Early examples of e-learning were the computer-based training applications often referred to as CBTs, mainly developed to foster professional (re)training. Then low-cost computers were developed, with reasonably good performance, and able to sustain multimedia applications and made available to the mass market. In more detail, it was the advent of the CD-ROM (mainly seen in this case as a mass storage device), along with easy access to the Internet that fostered the launch of the first mass-market-oriented CBT (mainly devoted to language teaching). Initially, products were conceived in a stand-alone fashion. Then hybrid versions (CD plus Internet) emerged to achieve a higher level of modularity and flexibility (including periodic product update), and now there are many online training courses.

According to IDC, e-learning means “educational content, learning services, and delivery solutions that support and enable Internet-based learning” (IDC, 2002). Thus, e-learning is a subset of what is defined as technology-based training (TBT) that includes CD-ROM and other technology-delivered training. In turn, TBT is a subset of what is called enhanced learning, which includes support to disability for learners, and so on. Finally, enhanced learning is a subset of the overall training and learning including instructor-led training (ILT) and text-based training, as shown in Figure 3, which elaborates on the basis of concepts exposed from the already quoted IDC report (IDC, 2002).
How Technology Can Support Culture and Learning

According to Tomita (1980) and Neuman (1991), the scenario presented in Figure 4 shows evidence of “a media gap.” Blogs, Wikis, virtual communities, and many other forms of communication and social/educational experiences are quickly arising and filling this “media gap.”

The availability of new and more pervasive technology support, as well as the growth and spread in access to culture and education, is ultimately addressing the issue of a “lifelong learning” process to avoid knowledge obsolescence, especially in environments like science or technology. We must bear in mind Novak (1998), when saying that the rote-meaningful learning continuum is distinct from the reception-discovery continuum for instruction (see Figure 5). As a matter of fact, it is not enough to have availability of information and technology that can assist in accessing the related knowledge.

A common mistaken notion is that “inquiry” studies will assure meaningful learning. The reality is that unless learners possess at least a rudimentary conceptual understanding of the phenomenon they are investigating, the activity may lead to little or no gain in their relevant knowledge. In this respect, it is worth noting that this aspect has been taken into account in the e-learning sectors and therefore, cognitive modelling and ergonomics have been closely investigated (Sussman & Sedivy, http://www.umass.edu/psychology/div2/eyelab/).

Studies on cognitive architecture have been carried out by taking into account perception, cognition, action (motor), and memory (Hornof & Halverson) as well as studies on eye movements and visual attention (Kowler). These results are extremely interesting and promising, as the use of these technologies could be very successful if employed to foster the learning process and support full accessibility, regardless of gender, race, language, and impairment. The introduction and combination of wireless communication, agents, adaptive user profiling, virtual reality, and eye-tracking (http://www.vr.clemson.edu/eyetracking/, http://www.nottingham.ac.uk/education/maps/eye.html) with artificial intelligence and other technologies can move the frontier for e-learning applications exactly in this direction.

The use of “multimedia” is probably one of the most relevant approaches to properly support education in all environments. It is an excellent way to support teaching activities, not only because it enables one to enrich the quality level of the service, but also because it allows one to attract the learner’s attention, and stimulates the mind from several perspectives. However, this is not enough unless the effort involved in overcoming accessibility problems and impairments in the learning process is taken into account. To achieve this result, the accessibility issue has to be addressed in the design phase and the available and emerging technologies, regardless of their origin and major field of application, have to be fully exploited. In other words, to examine carefully how accessibility-related needs, in general, and impairment could be tackled and use whatsoever technology that could fit to solve, or work-around, the problem so as to grant full access to education and culture and therefore, also greatly contributing to societal development.
Figure 4. The media gap (Adapted from Neuman, 1991)

Figure 5. Relation between learning autonomy and learner satisfaction in relation to media richness and bandwidth (Adapted from Novak, 1998)
As stated, in theory (as well as in practice), it is possible to devise usage of technologies presently applied in other sectors in the learning field to empower better and more compete forms of e-learning. As an example, just think how eye tracking has been exploited in the military environment to control firing systems⁵, and how it is presently used to support reading and comprehension exercises and other learning activities, but still well below its huge potential.

Another example may be the adoption of tools and methods used to teach diction, or opera singing, in the language learning environment (at least as far as pronunciation is concerned), especially when teaching languages like Chinese or Japanese to westerners or English to easterners. In this case there is a real need to learn things such as breath control, tone usage, phonation, and so on, just as it is done for opera singers or for actors.

**LANGUAGE LEARNING**

Most professional sectors have adopted technologies to support the learning process. Probably the first one to do so was the professional training sector (pilots, train conductors...) and language learning. Later came the informatics sector (programmers, EDP operators...), then the finance and banking sector (initially jointly with the informatics sector, then on its own, due to the increase in complexity of the environment itself), and lastly the medical training sector (initially in imaging and diagnostic to progressively shift towards a full coverage)⁶. This section focuses on the language-learning environment as an example, given that it represents a case where technology has already been widely adopted to properly support learning processes over a long period.

It is possible to find self-learning materials based on a widespread set of technologies (from vinyl audio support, to tapes, and now CDs or DVDs) and furthermore, many different learning methods have been adopted to support different classes of learners (businesspeople, primary and secondary students, home users, and so on). This illustrates the relation between e-learning and technology as a support for culture and education.

It is widely accepted that to master a language, the four linguistic skills (speaking, listening, reading, writing) have to be acquired, but the social and cultural background also have to be understood. Therefore, the study processes can be enhanced through the guided use of contemporary media such as newspapers and magazines, film and television, video, CD-ROMs, and Internet. Certainly, literature is an important part of the study of language and culture, especially at the more advanced levels, as it constitutes the most versatile and complex of language usages. Literature essentially represents a compendium of the past and present culture(s) of the language in which it is written. Therefore, it also represents the most difficult part to master when learning a foreign language as a second language.

When using a foreign language, words will turn out to be cages for concepts the learner wants to express. Unless the learner is natively bilingual, thinking directly in another language may be achieved only partially. For some people it is easier to learn the sound of the word/sentence and to note it down in their own language. This depends on many factors, starting from the degree of knowledge and mastery of the meanings associated with a word in a specific language, to end with the personal cultural background that may bias the nuances given to a specific word or expression (just think of all forms of professional jargon and disambiguation needs encountered when translating a word or a sentence out of its context). Many problems, misunderstandings, and issues have been caused by improper usage of words among people. Therefore, there should be a basic concern when teaching/learning a foreign language:
Concepts are expressed by words. If certain English words are missing in Chinese, it follows that the concept expressed by those words will be absent in China, and vice versa. (Mah, 2003)

When taking all this into account, it is evident that mastering a language is much more than learning its vocabulary and grammar; it implies also learning the culture behind the language; the history, the traditions, the costumes, and the related ways of thinking. In other words, all meanings, structures, metaphors, sounds, and tastes of a language need to be learned, in the same way a child from the country in question does this. This lesson has been used by scholars working in the language-learning domain, and is progressively being widely adopted in regular language learning in primary and secondary education. For example, as far as the “sound of a language” is concerned, exploiting songs, and their lyrics, has become a common way to support teaching, as is evident from the following quote by LeLoup and Ponterio (LeLoup & Ponterio, 2001):

Most foreign language teachers enjoy studying song lyrics as authentic text in their classes. Songs can be used at all levels and for a wide variety of activities and purposes such as comprehension, vocabulary introduction, illustration or recognition of grammar structures, and reinforcement of topics. Traditional or new children's songs, musical classics, or the latest pop hits are all fair game. The rhythm and melody of songs can make the words and expressions easier to remember and more enjoyable for students than other sorts of texts.

Unfortunately, teaching the culture behind a language is a long and difficult task, usually unfeasible in the timeframe of regular classroom teaching during primary and secondary education. However, this issue is being addressed more frequently than in the past, and is also often addressed by using e-learning in a lifelong learning approach. Furthermore, there exists class-based intercultural training devoted to specific audiences. Nowadays, learning words, sentences, and polite phrases in a foreign language, mainly to manage human/business relations when abroad, is becoming a common habit among far-sighted businesspeople, and therefore, it is essential to also teach proverbs, phrasal verbs, idioms, and so on. In the diplomatic environment, this has a long tradition; for example, in China, the first emperor studied Tibetan costumes and the language, to be able to properly welcome the Dalai Lama7 (Mah, 2003). Similarly, the Jesuit missionaries in Japan or China would spend long years in training before being sent to their final destination. During their training, they would learn how to behave, talk, act as if they were locals, and this training was finalised to achieve a specific purpose: the conversion of the locals to Christianity. In this, they reached such a level of perfection that Matteo Ricci is accounted among the classical Chinese authors in maths, geography, 8 and other topics, despite the fact he was an Italian.

The power of multimedia-based e-learning lies in the possibility to convey several media and all the related content richness to the user (in a single framework). For example, on a DVD, it is possible to convey as much data as in an ordinary encyclopaedia, with high-quality audio and good quality video. The Internet allows the updating of e-learning sources and tools in a combination of self and assisted teaching, making it possible to learn at one’s own pace and having tutor support only when needed. Synchronous activities can be performed along with asynchronous ones. Reinforcement tools are available, and all this is “virtually” usable by the learner.

There is a clear need to stress the word “virtually,” as disability support can be provided as well, but if this is lacking, then all this power may be lost for a relevant part of the population. From a technical point of view, it is possible to take ac-
cessibility (in the wider sense) into consideration from the very beginning, thus ensuring the widest possible coverage of user needs (from impairment to aging support). Nevertheless, this implies that proper development tools and strategies have to be used.

The outcomes from a 2001 European Commission open workshop to examine the future of e-learning expected the research focus to be on virtual and remote labs, emotional characters, handwriting recognition, and eye tracking in e-learning usability. All this should be combined with situational-based training and mobile access to resources so that: “Research on eLearning will support the development of innovative models for the provision of learning services, fully exploiting the potential of ambient intelligence technologies, enabling ubiquitous, interactive, personalized and tailored access to learning and knowledge on demand to individuals at home, at work or on the move. It will build advanced learning schemes for schools, universities, the workplace and for lifelong learning in general, reflecting the needs of the knowledge economy.” In other words, the pillars of the knowledge society and economy are ubiquitous access to personalized learning throughout life for everyone.

Traditional and Innovative Language Learning

Traditional class-based language courses are structured according to a well-codified approach that regards grammar and vocabulary to be the core of the study. The main differences that are worth noting in class-based learning are mainly related to target audience and objective of the class. Basically, we can distinguish the following categories of class-based learning:

- Basic institutional curricula (primary or secondary education level).
- Advanced institutional curricula (post secondary education level).
- Specifically oriented curricula (business, summer schools).

For each of these, there are specific programmes, schedules, and methods aimed to maximise the training outcome.

Except for literature or business-oriented courses, dialogues and drill downs are considered as complementary or support activities, just like exercises. Therefore, dialogues are usually linked to curiosities and hints on culture and civilization.

Figure 6. Language course levels according to several classifications

![Diagram of language course levels]

- Beginner – Berlitz 1 – A1 – ILR 1
- Beginner / Intermediate – Berlitz 2 – A2 – ILR 2
- Intermediate – Berlitz 3-4 – B1 – ILR 3
- Intermediate / Advanced – Berlitz 5-6 – B2 – ILR 4
- Advanced – Berlitz 7-8 – C1 – ILR 4
- Proficiency – Berlitz 9-10 – C2 – ILR 5

User proficiency
How Technology Can Support Culture and Learning

related to the language being studied. Classes are usually organized in terms of language knowledge prerequisites that are assessed via admission texts (this applies only to noninstitutional training). Yet even when assessment is not performed in advance, content and course structure reflect a similar approach with the presentation of the most difficult/uncommon concepts, authors, or issues at the most advanced level. A gradual approach to learning has always been adopted (see Figure 6), and is reflected in the courseware and content structure. This is certainly a harsh generalization, but still reflects a good part of the present class-based offers on the market.

The levels reported provide indications according to different classification approaches: UCLA, Berlitz (http://en.wikipedia.org/wiki/Berlitz_Language_Schools), common European framework of reference for languages (http://culture2.coe.int/portfolio/inc.asp?L=E&M=$t/208-1-0-1/documents_intro/common_framework.html), and Interagency language roundtable (http://www.utm.edu/staff/globeg/ilrhome.shtml, http://en.wikipedia.org/wiki/ILR_scale). Usually the passage from one level to the other is bound by the accomplishment of a specific (and in most cases well codified) exam. The schema also visualises the usual decline in participants from the basic to the more specialized training levels that are often not only more demanding in terms of prerequisites, but also in terms of competences and capacities of the trainees. Inside each course level, the typical learning unit could then be structured as follows (Figure 7).

This is a harsh schematisation, as the more advanced the course is, the less such a grammar-centric schema is followed, as there is a natural shift towards a more content-centric schema that is best tailored to the form of advanced learning provided. In fact, while the grammar-centric schema perfectly suits certain pedagogical purposes (like teaching how to teach a language), it is

---

**Figure 7. Typical structure of a grammar-centric learning unit**

<table>
<thead>
<tr>
<th>Level</th>
<th>Dialogue Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginner</td>
<td>Listen &amp; repeat (1)</td>
</tr>
<tr>
<td>Advanced</td>
<td>Listen &amp; complete (2)</td>
</tr>
</tbody>
</table>

(1) Here level is referred to the kind of exercises and not to the overall course that may be beginners, basic, intermediate, advanced, or proficiency.

(2) The student has to listen to a sentence and then repeat. There is not much interaction except repetition.

(3) The sentence is first heard, then repeated with blanks that the student should fill, and then repeated again.
not suitable for others, such as teaching literature or business/professional usage of a language. Obviously, a full range of intermediate models and combinations are also possible in order to accommodate the various needs and objectives of the training. A very similar approach has also characterised the evolution of e-learning-based language training content that was initially following the grammar-centric and is now shifting towards the content-centric unit. One of the main reasons for this shift also lies in the “blended” approach to teaching that combines traditional class-based to self-paced learning.

The difference in structure between the two approaches is quite apparent. Only a few products on the market are offering the user the chance to listen and modify voice samples or dialogues at playback speed. Some of the most effective online language courses, for Latin-based speakers willing to learn foreign languages, focus on reading/listening comprehension, grammar, and vocabulary. Users are provided with feedback on achieved performances, and exploit a pedagogical philosophy based on the following principles:

- The approach is student centred (user chooses own learning path).
- Each course level revolves around a central theme.
- Activities have a pedagogical rationale.
- The content is interesting and related activities are challenging.

Usually language training is divided into at least four different levels organized in about six courses. On average, each course is composed of more than 10 units plus a unit test. Units are, in turn organized, in several stimuli alternated with an average of five activities per stimulus plus tests. The course becomes progressively harder while progressing through units. For e-learning solutions foreseeing the adoption of CD/Internet-based content delivery, it is worth limiting the duration of each learning session to around 15-20

---

**Figure 8. Typical structure of a content-centric learning unit**

- **Dialogue**
  - Understanding & comprehension
  - Related grammar
  - Exercises

- **Drill down**
  - Culture
  - Society
  - Costumes & tradition
  - Slang & phrasal

- **Beginner**
  - Listen & repeat

- **Intermediate**
  - Listen
  - Listen & complete

- **Advanced**
  - Listen & complete
  - Listen

---

(1) A sentence is heard and repeated. Speed is partially adjustable.
(2) A sentence is heard as is and afterwards with blanks.
(3) Focus is on the usage and understanding of the language, including slang.
(4) A sentence is heard with blanks and the student should complete it; afterwards the sentence is heard as it should have been.
minutes for usability reasons, as schematically described in Figure 9.

On average, each stimulus is accompanied by a set of activities focused on different subskills, related to the various abilities, plus grammar and vocabulary. Listening is taught using slideshows, videos, or pictures with captions, while reading comprehension, and/or grammar are taught using texts.

Test units are presented at the end of a set of units aiming at providing feedback on what has been learned. Units may be repeated several times and should last no more than 30 minutes each. The structure will comprise a set of stimuli, each followed by a set of questions. At the end of the session, a score is provided as feedback. The score provided is in the format X out of Y correct answers (i.e., 3/5). If the same exercise is repeated several times, only the best score is reported. There may also be entry test units focused on comprehension and containing at least three graded texts. They should last no more than 20 minutes, be timed, and drawn from a bank containing several versions of the test (at least three). These entry units assign both a level and a course to the user upon completion, based on the decision schema reported in Figure 10.

According to the SCORM standard, a tracked course expects the user to complete the full unit before being granted the chance to proceed any further. Thus, if there are exercises for knowledge assessment, they can usually add conditions to the overall navigation procedure, for example, preventing the user from accessing any other section of the content until the present one has been completed and related tests passed positively.

---

Figure 9. Typical fruition scheme of a learning unit (innovative case)

Figure 10. Typical learner access level evaluation scheme of a language course
How Technology Can Support Culture and Learning

In general, content should be interesting and appealing. Tests should stimulate user curiosity and willingness to access the available content, providing exercises for the issues that generate most interest.

From the information presented, the structure of a typical learning course can easily be derived (see Figure 11). Please note that in most environments, the courseware structure is structured in levels according to the “European Council Curriculum” and to Flesch-Kincaid Grade Level Score, or to Flesch Readability Ease Score (at least as far as the English language is concerned).

Over time, several approaches to language teaching have been developed, starting from the traditional ones centred on grammar and stemming from the study of Latin and classical Greek. This evolution has been led by the following factors:

- Recognition that “live-languages” (currently spoken) differ from “dead-languages” and are constantly evolving; therefore, Understanding & Comprehension should come first, and the implied grammar could then be studied (and in some cases even derived).
- Spoken languages often rely on a smaller set of vocabulary, and the most common phrases structure is simpler. The usage of verb tenses and cases is also typically more limited in spoken language than in literary usage.
- Recognition of different needs for different audiences (business or technical trainee un-

Figure 11. Typical navigation scheme of an e-learning language course structured in levels
dertake a learning process to be able to communicate, usually in short timeframes).

- Provision of situational training better supports the retention of acquired knowledge than the extensive study of grammar and vocabulary.

A good example is the “Berlitz” method, or other derived ones like “inlingua” (http://www.inlingua.com/English/Services/main.asp?RefNav=rnav3). It is now clearer why there has been a progressive shift from a grammar-centric to a content-centric approach. This process greatly contributed to the need for effective and efficient self-learning that is now converging towards a “blended” approach, where the self-led activities can be even performed on the move, with the support of specifically designed tutoring assistance. This can be summed up in terms of distinctions applied to language learning depending on learning scope, the institution providing the training, trainees’ backgrounds, and domains, as illustrated in the Table 1.

### Language Labs

The setup of language labs dates back to at least the 1970s, when language teachers would use this infrastructure for practices and for assessing students’ progress, especially in two of the four abilities, namely listening and speaking, while reading and writing ability would be aligned with class work. The most frequent usage of the language lab would be either in “listening and comprehension” or in “listening and repeating.” Often (but this was much dependent on the teacher) in language labs, students would encounter audio sources that could span a wide range of topics, from read-aloud literature to radio programmes conversation, or even music (folk, traditional, pop).

In the late 1980s and early 1990s, the traditional language lab started to be sidelined by multimedia labs (initially devoted to purposes other than language learning). Only in very recent times has there been a convergence of language and multimedia labs into a single entity, a “multimedia language lab,” even though it is still quite popular.

<table>
<thead>
<tr>
<th>Environment</th>
<th>Provider</th>
<th>Learning objective</th>
<th>Target audience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic language learning</td>
<td>Primary and/or secondary education; Specific language schools.</td>
<td>Acquisition of the basic skills (read, write, listen, speak)</td>
<td>All categories of trainees; Businesspeople; Professionals.</td>
</tr>
<tr>
<td>Advanced learning education</td>
<td>Interpretation schools; Language universities; Specific language schools.</td>
<td>Linguistic refinement; Literature; Language teaching; Situational training.</td>
<td>All categories of trainees; University students; Businesspeople; Professionals.</td>
</tr>
<tr>
<td>Professional learning education</td>
<td>Interpretation schools; Language universities; Specific language schools.</td>
<td>Interpretation; Translation; Linguistic refinement; Literature; Language teaching; Situational training.</td>
<td>Interpreters; Translators; Dubbers; Businesspeople; Professionals (legal, science, technical or medical…).</td>
</tr>
</tbody>
</table>

Table 1. Distinctions applied to language learning
How Technology Can Support Culture and Learning

(especially for certain forms of studies) to retain the “classical” configuration.

Classical Language Lab Configuration

The classical configuration of a language lab is based on a set of audio equipment (tape recorders eventually combined with auxiliary inputs) usually divided into three categories: the teacher unit (the most richly featured and complete unit, comprising several inputs), the student unit (limited to a recording unit), and a set of supportive infrastructures (comprising the switch and control board available to the teacher). The control console and the switch and control board are physical, even though there are implementations using a computer-based approach (Figures 12, 16 and 17).

A classroom may contain only a few workstations, depending on the target usage, availability of room, complexity of cabling, and costs. The control unit available to the teacher is usually analogical (in the most advanced perhaps computer based) and so is the student unit. In the latter case, the schema becomes the one of Figure 13.

Multimedia Language Lab Configuration

The configuration of a multimedia language lab is based on a combination of computers and a set of equipment (including tape recorders combined with auxiliary inputs). Additionally, in this case, the whole setup is divided into three categories:

- The teacher unit
- The student unit
- A set of supportive infrastructures

While in the traditional language lab the control unit available to the teacher is usually analogical, in multimedia labs, the control unit is implicitly computer based. Moreover, in multimedia labs, it is often possible to find equipment and applications devoted to a wider set of usages.

Figure 12. Typical structure of a classical language lab configuration

Figure 13. Typical structure of a computer controlled language lab configuration
than language learning, as they were originally created to support ICT teaching and then extended to support language learning.

Presently, multimedia labs are also being provided with extensions enabling the usage of e-learning solutions. This allows for a varied set

**Figure 14. Typical structure of a basic multimedia language lab configuration**

![Basic Multimedia Language Lab Configuration](image1.png)

**Figure 15. Typical structure of an extended multimedia language lab configuration**

![Extended Multimedia Language Lab Configuration](image2.png)
of possible configurations and setups, like the following ones (see Figures 14, 15 and 16).

As is apparent from the previous schemas, the usual structure is based on a network connecting the individual positions, each equipped with the specific sets of supportive tools needed, comprising specific HW (recorder) and SW (applications). The class could be either fully physically bound (all work places are in the same physical location) or dispersed (not all work places are in the same physical location, but some are remotely connected); furthermore, the class could also be connected to external resources (like a VOD server, Internet, and so forth.). Audio/video conferencing equipment can also be used to support open and distance learning.

Additionally, in this case the overall arrangement will be then complemented with a proper layout that will provide two different sets of workstations for teachers and students. Again, a

Figure 16. Typical structure of a multimedia language lab class and teacher work place

Figure 17. Typical structure of a multimedia language lab teacher’s console
classroom may contain only a few workstations, depending on the target usage, availability of rooms, complexity of cabling, and costs of the overall setup, as in the following images.

Content Production

In the publishing environment, the content production chain follows well-codified and standardised processes that have been developing over several years. This process, and the actors involved, are schematically reported in Figure 18.

Cost control and overall IPR and copyright management are usually undertaken in a parallel stream, to ensure constant monitoring of high-risk factors. As far as the economic impacts are concerned, cost control has to monitor process development and ensure that it is kept in line with expectations and budgets, retaining its profitability (or even increasing it whenever possible) and therefore, is a good part of the standard management process.

On the other hand, IPR and copyright management interactions occur whenever an asset cannot be cleared and therefore, has to be replaced. This event may occur at any step, and the impact may be marginal or relevant, depending on a set of possible combinations of factors:

- The relevance for the publishing project of the object that could not be cleared.
- The reason for lack of clearance.
- The stage of the publishing process.
- The availability of a replacement/equivalent object, and so forth.

This refers to traditional, desktop and multimedia publishing as they present the highest number of contact points and overlaps.

A similar approach applies to the editorial process of any product, yet digital TV, serials, movies and in general products requiring video production are somehow different as the design and planning phase is much longer and has several by-products (storyboards, drawings, scripts, books...) that in certain cases will have their own existence and production cycle, which is reported in Figure 19 with the same approach of Figure 18.

Authoring Tools and E-Learning Content Development Applications

Standard document creation tools rank at the top of the list of utilities used to create e-learning content. Most of them are oriented to achieve specific results and do not require specific programming skills, even if they may have extensions, they are therefore often used in combination.

According to Brandon and Hall, over 40% of companies use an e-learning solution to provide services to their own personnel, including language learning, mainly due to globalisation. Most e-learning solutions adopted are used to provide self-paced e-learning courses and live e-learning sessions using virtual classroom applications, while tracking results using a learning management system (LMS).

Unlike conventional training, a good instructor cannot rescue bad online learning; the materials’ successful implementation is decided by the quality of the design. Therefore, in e-learning, content and design have remained the most critical dimensions of success. Not surprisingly, the shortage of well-designed, engaging, and relevant e-learning products is still high on the list of reasons for limited acceptance of e-learning in many organizations.

A Contextualised Example of Content Development for E-Learning

To provide a better understanding and to highlight the benefits of the adoption of certain technologies as a support tool for the production of educational content, we examine the process followed in a successful European RTD project conducted in IST-FP5 dealing with the application of technologies developed for the support of dyslexia to adult
### Figure 18. Content production chain in the publishing environment

<table>
<thead>
<tr>
<th>Activity</th>
<th>Task</th>
<th>Involved roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idea</td>
<td></td>
<td>Management</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Authors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chief Editor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Press Office</td>
</tr>
<tr>
<td>Market survey</td>
<td></td>
<td>Management</td>
</tr>
<tr>
<td>Title design</td>
<td></td>
<td>Management</td>
</tr>
<tr>
<td>Go / No Go decision based on market data and production cost analysis to ensure the expected return on investment</td>
<td></td>
<td>Management</td>
</tr>
<tr>
<td>Research of:</td>
<td>Sources, References, Contacts, Multimedia, Similar titles</td>
<td>Management</td>
</tr>
<tr>
<td></td>
<td>Authors, Editorial board, Editorial staff, Press Office, Legal Department</td>
<td></td>
</tr>
<tr>
<td>Draft acceptance</td>
<td>(if positive the next step starts if not the previous is reiterated)</td>
<td>Management</td>
</tr>
<tr>
<td></td>
<td>Authors, Chief Editor, Editorial board, Legal Department</td>
<td></td>
</tr>
<tr>
<td>Editing of:</td>
<td>Texts, Notes, Indexes, Multimedia, Captions</td>
<td>Authors</td>
</tr>
<tr>
<td></td>
<td>Chief Editor, Editorial board, Editorial staff, Instructional designer, Press Office, Legal Department</td>
<td></td>
</tr>
<tr>
<td>Product final acceptance</td>
<td>(if positive the next step starts if not the previous is reiterated)</td>
<td>Authors</td>
</tr>
<tr>
<td></td>
<td>Chief Editor, Legal Department</td>
<td></td>
</tr>
<tr>
<td>Finalisation of:</td>
<td>Texts, Notes, Indexes, Multimedia, Captions</td>
<td>Chief Editor</td>
</tr>
<tr>
<td></td>
<td>Editorial board, Editorial staff, Instructional designer, Production department, Press Office, Legal Department</td>
<td></td>
</tr>
<tr>
<td>Formal authorisation to start production</td>
<td>(if positive the next step starts if not the previous is reiterated. In this case IPR/© clearance should have been completed if not process may be suspended / stopped / cancelled)</td>
<td>Management</td>
</tr>
<tr>
<td></td>
<td>Chief Editor, Legal Department</td>
<td></td>
</tr>
<tr>
<td>Production</td>
<td>Books &amp; Magazines, CD-ROM, DVD, Web, TV, ITV, PDA, mobile and other new media</td>
<td>Production department</td>
</tr>
<tr>
<td></td>
<td>Outsourced service, Press Office, Marketing manager, Legal department, Company accountant</td>
<td></td>
</tr>
<tr>
<td>Marketing</td>
<td>Promoting</td>
<td>Marketing manager</td>
</tr>
<tr>
<td></td>
<td>Legal department</td>
<td></td>
</tr>
<tr>
<td>Distribution &amp; rights selling</td>
<td>Revenue management</td>
<td>Marketing manager</td>
</tr>
<tr>
<td></td>
<td>Legal department</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Company accountant</td>
<td></td>
</tr>
</tbody>
</table>
L2 language learning, named FLIC (Foreign Language Acquisition through the Instinct of a Child) (FLIC, 2004, 2005a, 2005b, 2006); (MediTECH, 2005a, 2005b); (Lasnier, 2005). The results of the project, in terms of pedagogical and neurological validation, have been conducted by Sheffield University, while the original idea and method at the basis of project development has been based on the experience gained by MediTECH in respect of the development of supporting tools for dyslexia treatments.

Given this content production process, it is apparent that producing content for multimedia applications, especially in the educational field for language learning, is a rather complex task per se; the introduction of new technologies that alter
such a process is often opposed due to the potential cost impact. FLIC proved that the additional effort required to take into account the specific solutions that will lead to product improvement is not excessive if this is done from the product design phase (FLIC, 2004, 2005a, 2005b); (MediTECH, 2005a); (Lasnier, 2005). For the sake of reference, we will briefly report here the process followed in FLIC.

In FLIC, the audio content was using dummy stereophony (see Figure 20) for recorded sounds. The dummy head simulates the human head in an acoustic sense, as far as possible, as in place of eardrums, the dummy head has suitable microphones whose directional characteristics and other physical properties correspond to the properties of human ears (FLIC, 2005a).

The need for this special recording procedure was strictly connected with the special learning support provided by the FLIC equipment, namely the lateralisation (MediTECH, 2005a), the multichannel voice fusion (HW/SW versions) (MediTECH, 2005a, 2005b), plus the time alignment and voice modification (SW version only).

The Lateralisation process is based on a quite simple principle related to the fact that specific aural circumstances generate both a higher level of attention in the listener, and also an instinctive synchronisation effect. This is achieved in three phases, namely:

- **Phase 1:** Learner hears model voice from CD and reads along silently phoneme by phoneme hearing the model voice “wandering from side to side.”
- **Phase 2:** Learner silently reads along with the model voice (mouthing the words), again perceiving it as “wandering from side to side.”
- **Phase 3:** Learner reads aloud in synchrony with the model voice, hearing his and the model voice in opposite ears and both “wandering from side to side”; this automatically tends to put the two in sync.

<table>
<thead>
<tr>
<th>Tool used to create e-learning courses content</th>
<th>Freq. of adoption</th>
</tr>
</thead>
<tbody>
<tr>
<td>PowerPoint / Word</td>
<td>66% / 63%</td>
</tr>
<tr>
<td>Dreamweaver (can be used by specialized and novice users)</td>
<td>61%</td>
</tr>
<tr>
<td>Flash (can be used by specialized and novice users)</td>
<td>47%</td>
</tr>
<tr>
<td>Code at the HTML tag level</td>
<td>34%</td>
</tr>
<tr>
<td>Traditional authoring tools</td>
<td>32%</td>
</tr>
<tr>
<td>Microsoft FrontPage</td>
<td>26%</td>
</tr>
<tr>
<td>Learning content management system (LCMS)</td>
<td>21%</td>
</tr>
<tr>
<td>Content authoring tools built-in to an LMS</td>
<td>18%</td>
</tr>
<tr>
<td>Rapid e-learning development tools</td>
<td>13%</td>
</tr>
<tr>
<td>Other</td>
<td>16%</td>
</tr>
</tbody>
</table>

Table 2. Documentation tools used with e-learning
The multi channel voice fusion (MCVF) technology divides the audio information of the left and right canal into eight (8) bands each, without losing any information during the splitting. The wave bands are differentiated in the following way: Band 1 (low-pass filter till 200Hz), Band 2 (250-315 Hz), Band 3 (400-630 Hz), Band 4 (800-1000 Hz), Band 5 (1250-1600 Hz), Band 6 (2000-2500 Hz), Band 7 (3150-4000 Hz), and Band 8 (high-pass filter from 5000 Hz). Consequently, two times 8 canals are generated. The mutual mixture of the canals causes an overlay of the separate audio information (left and right) to a single frequency composition of both canals. This principle is shown in Figure 22.

If MCVF and its effect are firstly used without lateralisation, while speaking through the microphone and listening to one’s own voice with and without the MCVF, it is impossible to locate the direction of the voice, which is precisely the desired effect, produced by the FLIC-unit for the benefit of the learner. Furthermore, a time-alignment procedure was also implemented (SW version only) to help the learner better understand the pronunciation aspects of the reference voice. The logic diagram of the adopted solution is shown.
How Technology Can Support Culture and Learning

It was mentioned that technology could support culture and learning by lowering access barriers. Technology has made access to content and

in Figure 23. Last, but not least, it was possible to modify the reference voice to enhance the possibility to understand mismatches occurring between the learner’s and the reference voice. This proved extremely useful (and was often needed) whenever the two voices belong to different genders. The result could be achieved with a voice modification engine (SW version only) that would operate according to the logics presented in Figure 24, transforming the original (reference voice) to achieve the result presented in Figure 25.

TECHNOLOGY AND ACCESSIBILITY

It was mentioned that technology could support culture and learning by lowering access barriers. Technology has made access to content and
sources mobile and ubiquitous, and it has also made it possible to overcome other barriers related to impairment. This is clear to a good part of the technical community, but not in totality, nor to the wider public. Such a lack of understanding of the importance of “accessibility from scratch” represents, at present, a real obstacle to societal evolution.

To get a clearer picture of the problem, it is sufficient to note that people with disabilities number nearly 1 billion worldwide, representing a formidable force of employees and customers. In Europe, according to the EU Commission, people with disabilities comprise about 15% of the population, and many of them face barriers when using ICT products and services.

Despite these data, many people still consider impairment to be a marginal problem and accessibility not a relevant issue. With aging, however, everyone starts experiencing (to some extent) impairments, and accessibility is a major issue.

The number of elderly people is actually rising fast, according to UN figures (18% of the population was aged over 60 in 1990; while 30% will be over 60 by 2030), and very often, elderly people encounter knowledge or usage barriers when using ICT products and services. This, therefore, brings back the issues of “lifelong learning” and “accessibility from scratch.”

Surprisingly, the technology required for achieving accessibility dovetails with technology required for pervasive computing, as nonimpaired customers need increasingly flexible, multimodal interfaces to accommodate IT access under varying environments, and with numerous devices. Therefore, taking impairment into account in the design phase of products and services automatically widens the market perspective and the societal benefit.

In 1998, the United States of America Congress amended the Rehabilitation Act to require Federal agencies to make their electronic and
information technology accessible to people with disabilities. This decision was based on the fact that inaccessible technology interferes with an individual’s ability to obtain and use information quickly and easily.

Section 508\(^{13}\) was enacted to eliminate barriers in information technology, to make available new opportunities for people with disabilities, and to encourage development of technologies that will help achieve these goals. The law applies to all Federal agencies when they develop, procure, maintain, or use electronic and information technology. Under Section 508 (29 U.S.C. ‘794d), agencies must give disabled employees and members of the public access to information that is comparable to the access available to others.

This process has brought about a major societal change in the USA and the upcoming spread of e-government solutions worldwide (http://www.w3.org/WAI/Policy/, http://www.w3.org/WAI/), and more specifically in Europe (Council Resolution of 25 March 2002 on the eEurope Action Plan 2002) means that the adoption, in the near future, of procedures and tools to make “full accessibility” a reality is now a necessity. E-learning will therefore be no more exception or something only for professionals (including or limited to training or retraining). Furthermore, the increase in the societal understanding of accessibility-related issues will favour a selection process where accessible products will be chosen over nonaccessible ones, even overriding the factors of pricing and brand name recognition.


All this leads back to the accessibility issue in the wider sense; as technology has reached a point where the objective could be reached to a great extent; what is lacking is both the application and regulation framework, along with a broad acceptance and recognition of this need for a real broad accessibility to content and education.

Previously, we have already provided a definition for e-learning and how it has come to the stage where, when talking about learning and accessibility, it is necessary to point out that e-learning is one of the terms that has emerged from the rapidly developing world of the Internet, and is broadly defined as “Internet-enabled learning.” This is usually referring to accessibility in terms of possibility to access, despite distance and time constraints, to sources, not in terms of impairment support. This is certainly not a minor issue.

Much expectation surrounds the Internet and its role in education, as e-learning can contribute to the improvement of standards and to the effectiveness of teaching and learning. Yet when claiming this, almost no reference is made to impairment support. Most Web-based content producers are often not even aware of W3C guidelines for accessible design of Web-based applications. The reason for this can probably be found in the fact that Internet access has generally only been available to society as a whole for the last 10-15 years, and to schools in the last 2-5 years. Such access is able to provide:

- Support for needs and individual learning requirements of both children and adults.
- A means of empowering children, helping them to learn more effectively.
How Technology Can Support Culture and Learning

- Access to vast volumes of information, mostly free.
- Communication tools and facilities that can be used on different levels.
- Access to content and learning processes that used to be confined to schools and colleges.

Summing up, the availability of Internet access provides a challenge to traditional forms of teaching and learning, and opens up opportunities previously denied to the majority. Yet all aforementioned factors are usually taken into account regardless of proper accessibility and usability related issues (in the sense that most content available on the Internet may fail either to meet accessibility issues in respect to regulations like the U.S. 508 one or similar, or usability criteria for certain user communities like elderly people or people with cognitive impairments).

This has happened largely due to the most widely adopted and common definition of e-learning. According to the IDC report “E-learning: The Definition, the Practice, and the Promise,” we can say that a good, working definition for e-learning (and in more general sense of e-anything) is “electronic” or “Internet-enabled.” Internet-enabled learning, or e-learning, strictly means learning activities on the Internet. Those events can be “live” learning that is led by an instructor or “self-paced” learning, where content and pace are determined by the individual learner.

The only two common elements in this process are a connection to the Internet (either physical or wireless) and learning. In 1996, a program for bringing technology into education was launched in the U.S. Its primary goals were:

- All students and teachers will have access to information technology in their classrooms, schools, communities, and homes.
- All teachers will use technology effectively to help students achieve high academic standards.

- All students will have technology and information literacy skills.
- Research and evaluation will improve the next generation of technology applications for teaching and learning.
- Digital content and networked applications will transform teaching and learning.

Presently, these objectives have almost been fully met in the U.S. Furthermore, according to the report “A Vision of eLearning for America’s Workforce,” developed by the Commission on Technology and Adult Learning, the global e-learning industry comprises approximately 5,000 suppliers offering every imaginable method of e-learning. The vast majority of these suppliers are private.

Even if it is generally accepted that the driving force behind the 21st-century economy has been knowledge, in many European countries, these objectives are still far from being realised.

What is driving the market in real time can be split among present and future trends, while drivers and inhibitors are illustrated in Tables 3 and 4.

A direct consequence of this is that technology needs to be used in order to lower costs. This can be achieved by reusing content (even if usually an increase of richness is also requested), and adding value by offering custom solutions and new approaches (performance support, customer education, and so forth.). Yet the most important issue regarding e-learning systems is related to its effectiveness in transferring knowledge to the user. Learners’ comprehension depends strongly on course structure. Therefore, particular attention must be paid to effective navigation and look-and-feel features. In addition, content should be designed, developed, or adapted for multimodal delivery for effective content reusability. Such integrated approaches are more likely to have an impact as the meta-adaptivity system design domain emerges.
How Technology Can Support Culture and Learning

It has been pointed out that at present, there is a shift in education that is moving from an approach in which individuals undergo a specific period of training and then start their professional life to an approach in which the learning process is a lifelong one. In this second perspective, companies also have to tackle the issue of continuous education.

For a rather evident set of constraints, including but not limited to time and costs, companies have been always trying to find alternative solutions to class-based learning. Thus, it is not surprising that some of the early adopters of e-learning have been major corporations like IBM, HP, AT&T, Boeing, and many others. In this respect, it is interesting to note that according to Brandon and Hall (Chapman, 2006; Chapman, 2006a Chapman & Nantel, 2006; Nantel & Vipond,

Table 3. Today and tomorrow (Source: Online Courseware Factory Ltd)

<table>
<thead>
<tr>
<th>Today</th>
<th>Tomorrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology training</td>
<td>Performance improvement</td>
</tr>
<tr>
<td>Classes for the masses</td>
<td>Personalised learning</td>
</tr>
<tr>
<td>Instructor centric</td>
<td>Learner centric</td>
</tr>
<tr>
<td>Training when scheduled</td>
<td>Learning on demand</td>
</tr>
<tr>
<td>Time to train</td>
<td>Time to perform</td>
</tr>
<tr>
<td>Teaching by telling</td>
<td>Learning by doing</td>
</tr>
<tr>
<td>Product-based learning</td>
<td>Project based learning</td>
</tr>
<tr>
<td>Know what</td>
<td>Know why</td>
</tr>
<tr>
<td>Skill and information (mastery basics = 3Rs)</td>
<td>Inquiry, discovery and knowledge basics</td>
</tr>
<tr>
<td>Reactive</td>
<td>Proactive</td>
</tr>
</tbody>
</table>

Table 4. Drivers and inhibitors (Source: Uwe Krueger, Realtech, http://www.realtech.de)

<table>
<thead>
<tr>
<th>Drivers</th>
<th>Inhibitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lifelong learning</td>
<td>Technology Compatibility</td>
</tr>
<tr>
<td>Economic turbulence</td>
<td>Limitations</td>
</tr>
<tr>
<td>Need for continued education and staff training</td>
<td>Billing systems</td>
</tr>
<tr>
<td>Increased Internet / eCommerce usage</td>
<td>Security concerns</td>
</tr>
<tr>
<td>Price</td>
<td></td>
</tr>
</tbody>
</table>
How Technology Can Support Culture and Learning

2006), “organizations’ eLearning practices are what turns tools and content into real learning.” Furthermore, to be successful in the design of e-Learning courses, the following factors should be taken into account:

1. **Faced problem and expected results:** Matching the e-learning solution to the problem at hand; achieving intended results.
2. **Instructional design and integrity:** Structuring, relevance and quality of content; focus on real-world competencies; selecting the right strategies for the content and context. This also means focussing on the intended users.
3. **Evaluation and assessment:** Applying imagination and rigour to the design and implementation of evaluation or assessment.
4. **Interactivity:** Using creativity and expert design practices to achieve instructionally powerful interactions of all kinds.
5. **Usability and interface:** Creating an effective, easy-to-use interface.
6. **Motivation and aesthetics:** Motivating learners to follow and successfully complete the training; hitting the right tone and aesthetic notes.
7. **Media and technology:** Smart selection and application of media, development tools and delivery technologies.
8. **Money and time:** Achieving excellence under constrained budgets and time lines.

In this process of integrating education into the employees’ daily life within a company, several aspects should be taken into account. For example, the need to keep certain components of company know-how inside the company, or how to foster practical skills acquisition, or how to handle the mapping of individual characteristics onto a professional job description to then derive, thanks to a gap analysis, the individual training needs.

Most of these items correspond in practice to one or more tool-sets that are used by the human resource department, the IT/ICT infrastructure, the employees, and the company management to address the various issues. The most relevant ones are described next, and could also be used in the “formal” education process to improve what is presently achievable, or to complement it in order to tackle specific issues.

**Authoring Tools, KMS and (L)CMS**

Authoring tools are generally desktop, single-user applications used to construct learning content by assembling and combining text, graphics, audio, video, and animations into e-learning courses. Yet there is also a current trend to have Web-based development tools. They have been initially derived by visual HTML editors and are often used to create standard Web pages, but they can also be used for creating learning applications. In addition, there exists a growing number of authoring tools focused on rapid, template-based development of learning content, sometimes with a specific focus on a particular type of learning application, such as software simulations.

Authoring tools, even though offering a rather broad range of solutions, represent only one category of tools aimed at properly editing and formatting content/knowledge. This is undoubtedly the first step in the process. Yet once content has been prepared it has to be managed, and a set of solutions has been developed for this purpose over time. Notably the most interesting are represented by:

- **Content management systems (CMS):**
  Systems for organizing and facilitating col-
laborative creation of documents and other content, using a range of technologies and techniques, including portal systems, wiki systems, and Web-based groupware; sometimes Web based, though in many cases, requiring special client software).

- **Knowledge management systems (KMS):** Systems for organizing and facilitating collaborative creation and management of knowledge and expertise, either at company or at individual level, using a range of technologies and techniques, including portal systems, wiki systems, and Web-based groupware.

- **Learning content management system (LCMS):** A solution for the creation, management, and transfer of learning content; it is apparent that LCMS are a subcategory of CMS devoted to learning content.

**E-Learning, ODL, and Lifelong Learning**

Computers are moving out of laboratories and into classrooms in learning new languages, understanding complicated math formulas, and exploring other countries. Technology is changing basic notions of schools and education and creating classrooms without walls that offer students a valuable view of the world, and enable them to experience and interact with other students and resources around the globe. With numerous forms of distance learning now available, public education is moving away from a need for students and teachers to assemble in school buildings for education to occur. This trend has far-reaching implications for the structure of state education budgets.

According to analysis performed by government and research institutions, the rapid growth in such areas as distance learning, technology-enabled assessment, and the increasingly diversified and expanded public-private learning marketplace require the development of new strategies for quality and consumers’ protection. Important priorities for the public and private sectors include:

- Providing reliable and universally accessible quality information for consumers.
- Developing quality assurance mechanisms.
- Ensuring that learners have the support they need to make the right decisions about their e-learning options.
- Developing policies and practices to ensure privacy.

Traditional, institution-based approaches to assessment and certification are not well suited to an e-learning world in which the focus turns from a record of classes taken and degrees received, to measures of what an individual actually knows and is able to do. As a result, private and public sector leaders need to take steps to create new approaches such as developing and promoting outcome-based assessments of learning results, and creating an electronic system for tracking those results.

Government and commerce must play a leadership role in making quality e-learning opportunities more widely available to all, from supporting the development of common technical standards to promoting broader access in under-served communities. The challenge and the opportunity are the same: to realize e-learning potential for reducing the divide between “haves” and “have nots.” Therefore, it is necessary to:

- Create the highest-quality e-learning experiences possible.
- Implement new measures and methods for assessing/certifying what users know and are able to do.
- Ensure broad and equal access to e-learning opportunities.

The potential return on investment for both the public and private sectors is enormous. The
challenge for businesses is to realize the full potential of e-learning as a driver of productivity and performance gains by making it an integral part of organizational strategy and operations. For government, the challenge is to create a nurturing policy environment for e-learning, firstly by removing barriers that restrict access to e-learning benefits and, secondly, by promoting industry self-regulation while balancing citizens’ interests and needs.

By adopting e-learning standards like the one defined by IEEE, AICC, and so forth, it is possible to achieve full interoperability of e-learning platforms and to have real, full portability of produced content. A learning object respecting all features described in those standards will be available in whatsoever platform and therefore, full independence from a specific vendor is achieved. Moreover, it is possible to package objects developed with different tools and for different delivery platforms, achieving a consistent and interoperable object, where the tracking will also be kept across editing, aggregation, and reuse. Other emerging standards that focus on personalisation are likely to have an influence in the near future, such as the internationalization of the IMS AccessForAll approach in ISO, the Individualized Adaptability and Accessibility in e-learning, Education and Training Standard.

Virtual Reality and Multisensorial Approach

The use of virtual reality is progressively becoming widespread. Usually the primary applications are either professional training or gaming. In the first case, the simulation is devoted to enable trainees to have access to equipment and situations that may be too risky and too costly to be faced otherwise. At the same time, only big organisations can presently afford costs related to simulations. On the other hand, namely in gaming, customers are demanding more and more in terms of appearance, and the huge numbers involved allow developers to achieve high levels of quality, despite the production costs.

With today’s technology, it is possible to achieve results unforeseeable just a few years ago. In detail, it is now possible to achieve multisensorial stimuli by combining visual, tactile, and sound experiences into a unique new one experience. The real challenge is to be able to generate similar situations (potentially involving even smell) for edutainment purposes. What is still lacking is a paradigm to convey all those rich stimuli into a single mainstream that may be used to both entertain and educate at the same time.

Virtual Communities and Forums

In many cases, establishing user communities is an extremely powerful and convenient way to foster a learning process. Community users will exchange ideas and cooperatively support each other in the learning process. This approach has a long history, and was initially born among scholars in the scientific domain as a way to support research results exchange, documents, and papers peer reviews, and other forms of interaction. With the wide diffusion of Internet, this approach has progressively widened its application, and presently there are thousands of communities, forums, and so forth. Given this, it is nevertheless essential to note that starting up a new community or forum is a difficult process. Table 5 reports the most relevant factors that would cause user retention or loss.

Table 5 reports the most relevant factors that would cause user retention or loss.

According to Table 5, actions that help in starting up the discussion in a new-born forum/community can be summarised as follows:

- Promote forum existence by direct contact with a set of people that may be interested to take part in forum activities and provide useful cooperation.
- Provide interesting threads.
- Periodically provide contributions from experts.
How Technology Can Support Culture and Learning

- Select key themes and collect them.
- Generate FAQs starting from relevant threads.
- Simplify the process to manage messages and replies, for example introducing the concept of a follow-up message that is not a reply but a related issue.

Another success factor that has to be taken into account is “need.” If the users needs to be part of a certain community, or if the users may find, in such a community, replies to their own needs, or even a sense of belonging/ownership, then community success is almost assured (the only real threat in that case may be economical sustainability). An interesting example for this is the German School Net; up to when it was strongly supported (in terms of costs and equipment) by Deutsche Telekom, it grew and developed. It turned out to be extremely important, full of users (both teachers and students) and data exchange. When Deutsche Telekom stopped supporting its activity, then community subsistence was greatly endangered. Now with the start-up of the European School Net, the German portal has also gained new vitality. Also of note is the case of eSchoolnet, the European Teachers’ portal.

### Table 5. Reasons for retention or loss within a new community

<table>
<thead>
<tr>
<th>Reasons to stay &amp; be active</th>
<th>Reasons to leave</th>
</tr>
</thead>
<tbody>
<tr>
<td>The forum is providing interesting news</td>
<td>Lack of a moderator</td>
</tr>
<tr>
<td>The content is complete and gives a good coverage of the topic addressed</td>
<td>A number of participants imposing their viewpoints</td>
</tr>
<tr>
<td>It is possible to find replies to questions or people able to provide such replies</td>
<td>It is not easy to understand the content and type of mail received from the forum</td>
</tr>
<tr>
<td>Participants are skilled, knowledgeable and interested</td>
<td>Too much mail is received</td>
</tr>
<tr>
<td>There is enough dialogue</td>
<td>Shift in focus</td>
</tr>
<tr>
<td>There is an historical track of performed discussions</td>
<td>Lack of education in participants’ behaviour</td>
</tr>
<tr>
<td>There is a service provision for FAQs</td>
<td></td>
</tr>
<tr>
<td>Case studies, white papers, reference documents, presentations, articles and papers are available (at least for browsing, better if downloadable)</td>
<td></td>
</tr>
</tbody>
</table>

### E-LEARNING AND THE MARKET

To complete the panorama provided so far to at least the relation presently existing between e-learning and the market should be addressed. According to IDC (2002), the e-learning market is segmented in three main areas: Content, Delivery solutions, and Services.

Content providers offer courseware core material (like multimedia assets and course structure), and testing/assessments are included in their expertise. Companies that provide delivery solutions offer products oriented at course preparation and organisation, such as training authoring tools and learning management systems (LMS), including collaborative software to help course fruition. Services include all additional features that can lead to better course preparation (content design), and that are related to course maintenance (management and hosting) and CRM (online mentoring). Some products/platforms provide solutions in different areas, covering partially or totally the entire range of e-learning offers. Figure 26 describes the situation and introduces market players and their
positions in the market as it appeared at the time of the previously mentioned IDC study.

**Asynchronous Web-based software suites** attempt to group suites of tools that quickly allow an instructor to convert existing basic electronic documents to a hierarchical system with relative ease. Although most of these suites may require limited HTML knowledge to fully exploit them, relatively little programming or other technical expertise should be needed to effectively use these software suites. Key characteristics of major players in asynchronous suites typically include capability for secure student login via standard Java browser, centralized database-
How Technology Can Support Culture and Learning

centred syllabus with links to internal or external Web pages, online, time-monitored quizzes with randomised dynamically-generated testing, discussion groups, and integrated e-mail. Systems also provide instructor development tools to ease transition from other media to these products. The academic user audience is dominated by two vendors: WebCT and Blackboard, with only a few additional competitors targeting academic audiences specifically. Corporate/government learning organizations have a much larger number of competing systems.

**Synchronous (real time) Web-based training solutions** are the most appropriate to facilitate relatively formal, instructor-led, hierarchical learning events, such as a seminar or interactive classroom. A number of these products incorporate "talking head video"; most video-enabled products suffer from poor video quality and tend to require 128KBPS (a.k.a. ISDN class) dial-in connections for acceptable performance; although a compelling sales feature tending to appeal to instructors familiar with videoconferencing, at present current "talking head video quality" over dial-up Internet connections is not worth the effort; most students ignore such poor images very early into the course. Leading members of this class generally have also the following characteristics:

- Browser-based (works with Netscape and Internet Explorer, though often PC-centric).
- Online testing (designed for use over Internet dial-up connections or corporate intranets).
- Live streaming audio (generally one-way, often two-way)
- Text chat, and occasionally private text chat, among students or between participants and presenter.
- Sequencing is controlled by a presenter-leader; a secondary/copresenter is available for the higher end products.
- Ability to show PowerPoint presentations.
- Ability to record sessions and make them available for on-demand viewing at any time.
- Shared whiteboard, with ability for students and faculty to import images that can be annotated on the whiteboard.
- Web-page sharing/cobrowsing.
- Application sharing ability for presenters to share applications running on their desktop.
- Integrated polling/surveys.
- Virtual “hand/raising,” to indicate when participants have questions.
- Firewall friendly.

**Application service providers (ASP)** choose to make their products available as hosted solutions only; they do not sell the software for organizations to create and run their own servers under license. This license arrangement may better fit specific needs, especially where institutions do not have the in-house capability to sustain tech support teams and servers. Vendors in this category generally offer the following feature set:

- **Browser-based:** All functionality is provided within a standard browser. Updates are added automatically.
- **Presentation sharing:** allows any meeting participant to spontaneously share any presentation (generally PowerPoint).
- **Document sharing:** allows meeting attendees to view content with multilevel zooming and annotation capabilities.
- **Application sharing:** Run any software application for effective live demos and training.
- **Desktop sharing:** Presenters can share anything on their PC system, including applications.
- **Web browser sharing:** Allows easy sharing of Web-based information.
How Technology Can Support Culture and Learning

- **Polling**: Lets presenters solicit quantitative feedback from attendees online.
- **Internationalization**: Ability to change the menus for selected languages.
- **Text chat**: Generally available as group chat and a separate channel for interaction during a session. A controversial feature in academics, as it is perceived by faculty as the electronic equivalent of note-passing or whispering in class.
- **Record and playback**: Permits recording of all interactions in a meeting for later reference, training, or demos. Anyone can play back a recording.
- **Application share control**: The presenter can share control of any software application with others in a meeting, for unmatched interactive meetings on the Web.
- **Scalable**: Supports multiple concurrent meetings with thousands of participants.
- **File transfer**: Users can conveniently upload and download files as desired.
- **Event tech support**: ASP vendor’s customer support personnel can take control of a user’s PC system (with the user’s approval) to instantly provide live assistance and resolve problems.

**Learning portals** are an emerging variation of Personal Collaborative Environments, and are an attempt to combine student services and community building through an integrated Web-enabled system, very much like the search engine portals Yahoo and Lycos. The use of the term “Learning portal” could imply either the basic software product needed to develop a portal or the actual use of such software to create a learning experience.

**Collaborative software**, an emerging category of software, allows individuals to interact one-to-one, peer-to-peer, or in small groups. The category may also be referred to as “Instant Messaging,” or “Buddy Systems,” although a number of products in this category also represent the older “listserve” e-mail-broadcast technology. This type of tool will become integral in advanced Web-based learning programs. Typically, tools in this category allow “awareness” of user-selected individuals and the ability to instant message (i.e., “chat” one-to-one) with individuals. Most programs provide for ad-hoc chatting in small groups. Some of these programs permit application sharing, voice-over-IP, or other useful over-the-Web features designed to allow individuals to interact. Currently, this category of software lacks interoperability. The Internet Engineering Task Force’s Instant Message and Presence Protocol represents the current form of standards proposed in order to solve these interoperability issues.

**Team groupware** is a new subset of collaborative software aimed at creating virtual project groups. “Team groupware” is distinguished from personal collaborative environments by these characteristics:

- Ability to develop documents synchronously and review asynchronously.
- One-to-one, one-to-many, and many-to-many capability.
- Centralized user management.
- Real-time Internet communications by at least one of the following:
  - Enterprise-ready systems scalable to hundreds of users.

Advanced software in this category includes group scheduling, ability to store and review older version of documents, and ability to route items. Systems should also require relatively little technical expertise for management after installation, as projects must be easy to create, edit, or terminate.

**Web-based education development tools** attempt to collect basic and single-purpose tools that meet specific needs but are not themselves intended to be a turnkey implementation sys-
How Technology Can Support Culture and Learning

tem. They are generally visual authoring tools or programmers toolkits, or meet a very specific Web-based need. This category has undergone substantial consolidation since the high-flying days prior to the dot.com crash in 2000/2001. A late 2002 study from Market Data Retrieval revealed that one-third of K-12 schools are already offering a number of distance learning programs for their students. High schools, in particular, have been attracted to the virtual classroom environment, with 53% offering such programs to students in rural states where students do not have access to specialized teachers. Furthermore, 36% of schools reported to have distance learning programs for their teachers, allowing them to access professional development programs, which is similar to the applications corporations are finding for e-learning.

REFERENCES


CETIS, the Centre for Educational Technology Interoperability Standards. Retrieved from http://www.cetis.ac.uk/


FLIC Project. (2005a). *D2.4 Voice models recording*.

FLIC Project. (2005b). *D2.3 Development of materials for the evaluation*.

FLIC project. (2006). *Special project status report for the EC*.


How Technology Can Support Culture and Learning


ENDNOTES

1 The term “clerices” survives in some students’ organisations, named “Goliardia,” that can still be found in ancient European universities.

2 St. Edmund Hall (1278), Exeter (1314), Oriel (1326), Queen’s (1340), New (1379), Lincoln (1427), All Souls (1438), Magdalen (1458), Brasenose (1509), Corpus Christi (1517), Christ Church (1546), Trinity (1554), St John’s (1555), and Jesus (1571).

3 Africans (RSA and South African Radio Broadcasting Corporation), Spanish (Radio Andorra), Chinese (Radio Beijing), Jew (Israel Broadcasting Authority), French (Radio France International), Japanese (Radio Japan), Greek (Cyprus Broadcasting Corporation), English (BBC, Voice of America in Special English, Radio Australia, KGEI California, WYFR), Dutch (Radio Nederland), Polish (Radio Poland), Russian (Radio Moscow), Swedish (Radio Sweden), German (Deutsche Welle, Deutschlandfunk, Radio DDR).

4 Originally named APPANET (from the combination of the names ARPA - Advanced Research Projects Agency and Net) was born during the late 1960s as a Defense Advanced Research Projects (DARPA) project to ensure efficient and robust communication among military bases and the capital, in case of a nuclear attack.

5 Some of the most well-known systems are the one used in avionics (e.g., on the Cobra attack helicopter), or in tanks and many other combat systems.

6 This is just a rough description of evolution partition and sequencing in the field
Actually, the first emperor ordered that a 1:1 replica of the Dalai Lama palace was built so that his guest could feel at home.

Matteo Ricci is the author of the first world map printed in China, which notably is China-centred, well in accordance with the Chinese definition of their land: “All that is under Heaven.”

Initially based on a combination of books and records, then tapes, and now CD/DVD

NotePad or other text editor

Director, Authorware, ToolBook, and so forth.

ReadyGo, Lectora Publisher, Trainersoft, and so forth.

http://www.section508.gov/index.cfm


This work was previously published in Interactive Multimedia Music Technologies, edited by K. Ng and P. Nesi, pp. 195-230, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 3.25
Tangible User Interfaces as Mediating Tools within Adaptive Educational Environments

Daria Loi
RMIT University, Australia

ABSTRACT

This chapter proposes that, as approaches to human computer interaction (HCI), tangible user interfaces (TUIs) can scaffold rich classroom experiences if they are coupled and generated within multi-pedagogical frameworks that adopt concepts such as Multimodality, Multisensoriality, and Multiliteracies. It overviews some necessary conditions for these tools to be effective, arguing that tangible user interfaces and multi-pedagogies are efficient when they are conceptualized as part of adaptive educational environments—teaching and learning ecologies where learners and teachers are seen as co-creators of content and of new ways of interacting with such content.

INTRODUCTION

Information and communications technologies (ICTs) enable types of learning experiences involving HCI that other means do not easily achieve (Simon, 1987; Underwood & Underwood, 1990). While digital spaces are traditionally manipulated via simple input devices (e.g., keyboard and mouse) that are used to manipulate representations displayed on output devices (e.g., monitors), tangible user interfaces remove this input-output distinction and connect physical and digital worlds using physical objects as interfaces to digital information (Ullmer & Ishii, 1997).

This chapter discusses the potential role of tangible user interfaces in scaffolding rich HCI classroom experiences and some necessary conditions for such tools to be effective. I argue that these interfaces can have a key role in contemporary teaching and learning environments if they are coupled and generated within multi-pedagogical frameworks that adopt concepts such as Multimodality (Kress & Van Leeuwen, 2001), Multisensoriality (Ceppi & Zini, 1998), and Multiliteracies (Cope, Kalantzis, & New London Group, 2000).
Tangible user interfaces and multi-pedagogies are, however, effective when they are conceptualized as part of adaptive educational environments (Loi & Dillon, 2006)—teaching and learning ecologies where learners and teachers are seen as co-creators of content and of new ways of interacting with such content.

This chapter is divided into four sections. In the first part I overview why and how tangible user interfaces can enrich classroom experiences, while in the second I outline the importance of coupling them with multi-pedagogical frameworks. The third section overviews the notions of adaptive educational environments and proposes that tangible user interfaces can be conceptualized as mediating tools enabling a shift of such environments to become creative spaces. In the last section I offer a number of concluding remarks, highlighting future implications and the need for new ways of conceptualizing contemporary learning environments.

One of the key objectives of this chapter is to highlight the importance of designing tangible user interfaces for teaching and learning by considering them part of larger ecological HCI frameworks where pedagogy, people, and context play a crucial role.

TANGIBLE USER INTERFACES AND LEARNING ENVIRONMENTS

Ullmer and Ishii (1997) point out that while graphical user interfaces (GUIs) have proven to be “a successful and durable model for human computer interaction,” the GUI approach to HCI “falls short in many respects, particularly in embracing the rich interface modalities between people and the physical environments they inhabit” (p. 1). Within this context, a range of alternatives has been explored, from ubiquitous computing to augmented reality. However, these attempts often rely on exporting the GUIs paradigm to world-situated devices, failing in capturing the richness of physical-space interactions they want to enhance (Ullmer & Ishii, 1997).

This understanding was the basis on which the notion of tangible user interfaces—user interfaces that adopt surfaces, instruments, physical objects, and spaces as physical interfaces to digital information—was constructed through initial explorations by Fitzmaurice, Ishii, and Buxton (1995) and the original work of Ullmer and Ishii (1997). Tangible interfaces put emphasis on touch and physicality in both input and output and are often coupled to physical representations of actual objects (O’Malley & Stanton Fraser, 2004). Examples of tangible user interfaces include rehabilitation tools (Edmans, Gladman, Walker, Sunderland, Porter, & Stanton Fraser, 2004), drawing and designing tools (Ryokai, Marti, & Ishii, 2004), collaborative and management tools, browsers and exploratory tools (Piper, Ratti, & Ishii, 2002), multimodal interactive tools (Raffle, Joachim, & Tichenor, 2002), music creation and performance (Patten, Recht, & Ishii, 2002; Weinberg & Gan, 2001), and toys/educational tools (Mazalek, Wood, & Ishii, 2001; Vaucelle & Jehan, 2002).

O’Malley and Stanton Fraser (2004) stress the beneficial role of physical manipulatives in learning environments by highlighting that:

• physical action and concrete objects are important in learning;
• physical materials trigger mental images that inform future problem solving in the absence of physical materials;
• learners can abstract symbolic relations from a variety of concrete instances; and
• familiar physical objects are more easily understood by children if compared with more symbolic entities.

This suggests that children are capable of demonstrating their knowledge via physical actions (e.g., gestures) and can solve problems by working with given concrete materials even if they cannot
solve them symbolically or in their heads. This is confirmed through the work of Bruner (1966), Piaget (1953), Church and Goldin-Meadow (1986), Goldin-Meadow (2003), and more recently Martin and Schwartz (2005).

Many thinkers have developed tools/environments that recognize the significance of children’s interactions with the physical environment: Pestalozzi (1894) and his method; Dewey (1938) and his philosophy of experience and relation to education; Montessori (1917, 1949) and her method; Kamii & DeVries (1978, 1980) and their take on Piaget’s theory; and Gardner and Hatch (1989) and their multiple intelligences.

The role of physical materials in triggering mental images useful for problem-solving has been discussed by several authors and teachers, including: Montessori (1917); Dienes (1964); Bruner (1966); Stigler (1984); Uttal, Scudder, and DeLoache (1997); and Chao, Stigler, and Woodward (2000).

The capacity to abstract symbolic relations from concrete instances is discussed by Bruner (1966) when he talks about enactive, iconic, and symbolic forms of representation and by Karmiloff-Smith (1992) and her theory of representational re-description where children’s representations move from implicit to more explicit forms.

In synchrony with the previous body of literature, tangible user interfaces and manipulatives have a number of characteristics (O’Malley & Stanton Fraser, 2004) including:

- they allow for parallel input (e.g., two hands) improving the expressiveness or the communication capacity with the computer;
- they take advantage of well developed motor skills for physical object manipulations and spatial reasoning;
- they externalize traditionally internal computer representations;
- they afford multi-person, collaborative use;
- physical representations embody a greater variety of mechanisms for interactive control;
- physical representations are perceptually coupled to actively mediated digital representations; and
- the physical state of the tangible embodies key aspects of the digital state of the system.

There are a number of examples of tangible user interfaces that enrich HCI in classroom experience through these characteristics. O’Malley and Stanton Fraser (2004) have reviewed several of such examples, dividing them into three main categories.

The first category is that of digitally augmented paper and books, such as Listen Reader (Back, Cohen, Gold, Harrison, & Minneman, 2001), MagicBook (Billinghurst & Kato, 2002), the KidStory project (Stanton, Bayon, Neale, Benford, Cobb, Ingram, et al., 2001), which adopts the KidPad storytelling software (Druin, Stewart, Proft, Bederson, & Hollan, 1997), and LeapPad®. Another category of tangible user interfaces is that represented by phicons (physical objects—such as toys, blocks, and physical tags—used to trigger digital effects), such as the CACHET project (Luckin, Connolly, Plowman, & Airey, 2003), Storymat (Ryokai & Cassell, 1999), the Tangible Viewpoints system (Mazalek, Davenport, & Ishii, 2002), and Chromarium (Rogers, Scaife, Gabrielli, Smith, & Harris, 2002). There are also digital devices with embedded computational properties, such as Triangles (Gorbet, Orth, & Ishii, 1998), Curlybot (Frei, Su, Mikhak, & Ishii, 2000), Topobo (Raffle, Parkes, & Ishii, 2004), Thinking Tags (Borovoy, McDonald, Martin, & Resnick, 1996), and sensors and digital probes (tangible devices based on physical tools that act as sensors or probes of the environment), such as Storytent (Green, Schnädelbach, Koleva, Benford, Pridmore, & Medina, 2002), I/O Brush (Ryokai et al., 2004),
Tangible User Interfaces as Mediating Tools within Adaptive Educational Environments

and the SENSE project (Tallyn, Stanton, Benford, Rowland, Kirk, Paxton, et al., 2004).

Tangible user interfaces have a beneficial role in learning environments due to their physical and tangible characteristics and trigger HCI learning experiences that can have different qualities from those in which a learner can engage by using other digital resources such as the Web. Moreover, these tools can enable many different ways of integrating face-to-face and virtual learning, creating hybrid interactions via their tangibility and manipulability nature. However, tangible user interfaces cannot operate and should not be developed in isolation from other key factors, as discussed in the next section.

TANGIBLE USER INTERFACES AND MULTI-PEDAGOGIES

As mentioned, tangible user interfaces enable numerous ways of integrating face-to-face and virtual learning. However, I propose that to be effective in the classroom and to enhance rich HCI learning experiences, these media devices need to be coupled with multi-pedagogical frameworks. Failing to do so implies the development of technology for technology sake instead of rich ICT tools that can enhance, support, and scaffold rich HCI learning experiences. As Reimann and Goodyear (2004) highlight, “technology is not so important in itself: what matters is how the technology is used” (p. 2)—pedagogy and pedagogical choices in making good use of technology have crucial roles.

I define multi-pedagogical frameworks as scaffolds that refer to and integrate notions of Multiliteracies (Cope et al., 2000), Multimodality (Jewitt & Kress, 2003; Kress & Van Leeuwen, 2001), and Multisensoriality (Ceppi & Zini, 1998).

The term multiliteracies refers to the variability of meaning making in different cultural, social, or professional contexts and to the nature of new communications technologies (Cope & Kalantzis, 2005). Cope et al. (2000), drawing from a number of theories, generated the Multiliteracies model by merging four approaches to teaching and learning: situated practice, overt instruction, critical framing, and transformed practice. Deepening the notion of multiliteracies, Kalantzis and Cope (2003) look at pedagogy in terms of eight knowledge processes, which are selected and deployed by the teacher and define the mindful and premeditated use of such processes as learning-by-design (Cope, Kalantzis, & the Learning by Design project group, 2005). In the Learning-by-Design project, “the teacher becomes a reflective designer of learning experiences (teacher-as-designer), and classroom plans become shareable designs-for-learning” (Burrows & Loi, 2006, p. 2). In this model meaning is generated in the space where visual, audio, gestural, and spatial patterns of meaning interface written-linguistic modes of meaning (Cope & Kalantzis, 2005). To Kress (1999) the visual “is becoming more prominent in many domains of public communication” and is “taking over many of the functions of written language” while displacing the written language “from its hitherto unchallenged central position in the semiotic landscape” (p. 68).

Building on the preceding considerations, the notion of multimodality implies a paradigm shift in how the communication and representation of meanings can be interpreted. A multimodal text adopts several modes of communication (such as image, writing, and speech), and, while multimodality characterizes many modern texts, “in the age of digitisation, the different modes have become the same at some level of representation, and they can be operated by one multi-skilled person” (Kress & Van Leeuwen, 2001, p. 2). The notion of multimodality has been to date discussed by a number of authors from various disciplinary domains. Besides Kress and Van Leeuwen (2001), the works of Norris (2004), Levine and Scol-
Tangible User Interfaces as Mediating Tools within Adaptive Educational Environments

lon (2004), Kramsch, A’Ness, and Lam (2000), and Jewitt (Jewitt & Kress, 2003) represent key examples.

Ceppi and Zini (1998), when discussing the characteristics that learning environments should have, talk about multisensoriality and multisensorial spaces: complex environments “made up of sensory contrasts and overlappings that are phenomenologically distinct” (p. 17). To the author, the space should not simply be about “being simply rich in stimuli,” but it should have “different sensory values so that each individual can tune into his or her own personal reception characteristics,” implying that “standard univocal solutions cannot be conceived” (Ceppi & Zini, 1998, p. 17).

In multi-pedagogical frameworks teachers deploy a number of knowledge processes in a premeditated and mindful way (multiliteracies), becoming reflective designers of learning experiences where meaning is represented and unfolded in multimodal ways (multimodality) and enriched by different sensory values (multisensoriality). As already highlighted, pedagogy and pedagogical choices have crucial roles in making good use of technology in the classroom, and technology for the classroom should be conceived and developed in light of and interlinked with such pedagogical choices. Within this context, this chapter argues that the design and development of tangible user interfaces for the classroom should be coupled with multi-pedagogical frameworks.

TANGIBLE USER INTERFACES AND LEARNING ECOLOGIES

Tangible user interfaces have a key role in scaffolding rich HCI classroom experiences, and they should be coupled with multi-pedagogical frameworks. To amplify these notions, I suggest that tangible user interfaces and multi-pedagogical frameworks are in fact part of learning ecologies, together with learners, teachers, and context, and, of course, the relationships among all actors (refer to Figure 1). These five actors are interrelated similarly to how context, people, practices, and relationship interact within collaborative workspaces (Loi, 2003). Reggio Emilia schools offer an exemplary instance of how HCI can be part of the learning ecologies (Project Zero & Reggio Children, 2001).

This model of learning ecology describes a situation where actors are deeply interrelated and co-dependent—there is a sense of wholeness and rich complexity. This implies that tangible user interfaces cannot be developed in isolation from other key actors, but rather conceived in light of their users (teachers and learners), contexts of HCI use, and designed learning experiences (multi-pedagogical choices).

Previous work (Loi & Dillon, 2006) discussed how ideas about collaborative workspaces (Loi, 2005) and integrativism (Dillon, 2006) can be incorporated and extended to look at the ecological nature of learning environments and to unfold notions of adaptive educational environments—spaces that accommodate changing relationships between people and resources—and creative spaces—environments that have been modified through designed interventions. Creative spaces
I suggest that tangible user interfaces could be employed as HCI mediating tools within learning ecologies that are conceptualized as adaptive educational environments. This implies that the dynamic flow underpinning the previous model (in Figure 1) requires a refinement (refer to Figure 2). This model maintains the ecological nature of Figure 1, while underlining the possibility to consider tangible user interfaces as mediating HCI agents (or brokers). This role and its positioning within educational environments impact how tangible user interfaces can be conceived, developed, and adopted. In line with Loi and Dillon (2006), the notion of tangible user interfaces as HCI mediating tools can be extended to re-conceptualize them as designed interventions that can trigger the development of creative spaces. These concepts will be explored in future work.

To address the development of tangible user interfaces in the described HCI context, I suggest that a participatory design approach (Sanoff, 1990; Schuler & Namioka, 1993) should be adopted to ensure that all parts of the whole are active participants in the design process. This proposition implies that teachers and learners must be involved in the design process; designers and technologists must operate in multidisciplinary teams together with teachers and pedagogy experts; and the contexts where technology will be adopted/embedded must be considered at all times. There are a number of examples of teams currently developing tangible user interfaces within these participatory paradigms. An interesting example within the educational context is offered by the Children as Design Partners project.

As previously mentioned, Reimann and Goodyear (2004) propose that in educational environments, “technology is not so important in itself: what matters is how the technology is used” (p. 2). Their statement, which implies that pedagogical choices in making good use of ICT have crucial roles, should be amplified because when considering HCI learning ecologies, the design of ICT has an equally crucial role, as it strongly impacts what pedagogical choices can be conceived and enacted. The interrelations and co-dependencies of technology and pedagogy should not be underestimated. In light of this, I suggest that tangible user interfaces can offer significant opportunities for effective use of HCI within the classroom if:

- they are conceptualized as actors within learning ecologies, together with multi-
Tangible User Interfaces as Mediating Tools within Adaptive Educational Environments

- there is a strong appreciation of the deep interrelationships and co-dependencies existing among learning ecologies’ actors;
- they are conceived as mediating tools within learning ecologies;
- the learning experiences that are mediated by tangible user interfaces are designed adopting multi-pedagogical frameworks; and
- they are developed adopting a participatory design approach.

The adoption of a participatory design approach can be extended to the development of rich HCI learning experiences. When looked at through such lenses, teaching and learning practices can be expanded to notions of teachers-creators, learners-creators, and teacher-student as space for content emergence (O’Malley & Stanton Fraser, 2004); teachers as author-publishers and project-consultants (Burrows, 2005a, 2005b); teacher-as-designer (Brown & Edelson, 2003); and teachers and learners as co-researchers (Loughran, Mitchell, & Mitchell, 2002; Phelps, Graham, et al., 2004).

Tangible user interfaces offer some significant benefits to classroom activities, in particular from (1) creativity, (2) diversity, and (3) collaborative perspectives. These are discussed in the following paragraphs.

Creativity is an “essential life skill, which needs to be fostered by the education system(s) from the early years onward” (Craft, 1999, p. 137). In her literature review on creativity, technology and learning, Loveless (2003, p. 13) overviews the characteristics of learning environments that are conducive to creative developments, including opportunities for exploration and play with materials, information and ideas (Craft, 2000), flexibility in time and space for the different stages of creative activity (Claxton, 1999), and opportunities to take risks and make mistakes in a non-threatening atmosphere (Davies, 1999). These are in synchrony with some characteristics typical of tangible user interfaces, such as their physicality, manipulability, concreteness, and familiarity (O’Malley & Stanton Fraser, 2004).

Tangible HCI enables learners to combine and recombine the known and familiar in new and unfamiliar ways (Hoyles & Noss, 1999) and to unfold the world through discovery and participation (Soloway, Guzdial, & Hay, 1994; Tapscott, 1998). As Hoyles and Noss (1999) stress, “it is this dialectic between known and unknown, familiar and novel, that provides a motor for creativity” (p. 19). Moreover, combining familiarity with unfamiliarity can promote reflexiveness (Rogers et al., 2002; Scaife, 2002), stimulating awareness and enhancing learning (Ackerman, 1996; Piaget & Inhelder, 1967).

Green, Facer, Rudd, Dillon, and Humphreys (2005) have suggested that as mobile, tangible, and networked technologies are increasingly deployed in facilitating creative work, educational environments will be characterized by a high degree of personalization where individuals will have greater influence over how they choose and use learning resources. The education system is constantly challenged by diversity issues, and researchers/educators are regularly looking for ways to enable learning for diverse learners (refer, for instance, to Cope et al., 2000). Through his work on the Internet generation, Prensky (2001) argues there is a need to radically rethink teaching practice to mirror current ways of learning. The diversity issue implies a need to change the education system so that it can conform to learners. Interestingly, learners are already shaping their own learning outside the classroom, using digital resources to create personalized learning environments for themselves outside of school. As a matter of fact, it has been suggested that by the age of 21 the average person will have spent 15,000 hours in formal education, 20,000 hours in front of the TV, and 50,000 hours in front of a computer.
Tangible User Interfaces as Mediating Tools within Adaptive Educational Environments

screen (International Futures Forum, Futuribles, European Journal of Education, & Université de Paris Dauphine, 2005). Various studies provide practical example of this tendency (for instance, Ba, Tally, & Tsikalas, 2002; Marsh, 2004).

The preceding data reiterates the need to find ways to shift the educational system so that it can conform to learners, better addressing their diversities. As Green et al. (2005) observe, the logic of education systems should be reversed so that it is the system that conforms to the learner, rather than the learner to the system. Thanks to tangible user interfaces, their intrinsic characteristics (physicality and active-oriented nature; sensory engagement; accessibility; concreteness and stimuli to abstract symbolic relations; capacity to trigger mental images that inform future problem-solving; familiarity) could offer interesting possibilities to this needed shift, enabling interactive HCI classroom activities where diverse learners can be in charge of how they choose and employ available learning resources.

A third significant benefit to classroom activities offered by tangible user interfaces is represented by their capacity to encourage collaboration as they afford multi-person, collaborative use (O’Malley & Stanton Fraser, 2004). A number of researchers have discussed and tested the benefits of: children working together (Rogoff, 1990; Stanton et al., 2001; Wood & O’Malley, 1996); collaborative practices on learning (for instance, Brandt, 1990; Hurley, Boykin, & Allen, 2005; Johnson, 1994; Johnson, Johnson, & Scott, 1978; Johnson, Maruyama, Johnson, Nelson, & Skon, 1981; Slavin, 1987); information technologies in peer and cooperative learning (for instance, Anderson, 2005; Chen & Morin, 2006; Topping, 2005); and that of tangible media in collaborative classroom practice (Africano, Berg, Lindbergh, Lundholm, Nilbrink, & Persson, 2004).

CONCLUSION AND FUTURE WORK

Although ICTs are already integrated into teaching across the curriculum, more encouragement might be given to teachers in terms of incorporating everyday technologies into the classroom, and the development of physical and multisensory activities, increasingly part of the curriculum, should occur with regard to the use of ICT (O’Malley & Stanton Fraser, 2004). In particular, tangible user interfaces could be considered and designed as mediating agents between pedagogy, teachers, learners, and context. This can occur only if more emphasis is placed on learners’ practices and teachers’ pedagogical choices and if teachers and learners are co-designers in the development of tangible HCI for learning.

There is a need to expand the ways in which tangible HCIs are researched, developed, deployed, evaluated, and implemented, and it is necessary to bridge the gap between technologist, designers, researchers, and teachers as their technical, pedagogical, and methodological knowledge and skills differ while complementing each other. A way to translate, mediate, and share ways of doing, thinking, and creating effective HCI must be developed.

As O’Malley and Stanton Fraser (2004) stress:

More research is needed on the benefits of tangibles for learning—so far the evaluation of these technologies has been rather scarce. More effort is also needed to translate some of the research prototypes into technologies and toolkits that teachers can use without too much technical knowledge. Teachers also need training in the use of non-traditional forms of ICT and how to incorporate them into teaching across the curriculum. And finally, new forms of assessment are needed to reflect the potential of these new technologies for learning.
The notion of **tangible user interfaces** as mediating agents within educational environments implies a shift in how the learning process is viewed, a shift in the roles of teachers and learners, and a shift in how learning spaces are shaped and interacted with.

O’Malley and Stanton Fraser (2004) express their hope that teachers will **take inspiration from the whole idea of technology-enhanced learning moving beyond the desktop or classroom computer by, for example, making links between ICT-based activities and other more physical activities.** I share this hope, and this chapter aims at nurturing this notion by linking technological aspects of HCI with pedagogical, ecological, and methodological considerations under the same holistic framework. The proposed HCI scaffold fosters a different way of looking at technology-design and technology-deployment, while promoting the opportunity to re-conceptualize roles, practices, and the relationships among key actors.

As discussed in this chapter, **tangible user interfaces** offer some significant benefits to classroom activities that involve HCI, in particular from creativity, diversity, and collaborative perspectives. However, these benefits imply a need for educational institutions to re-consider and re-address how they view the professional development of teachers, the classroom space, and how curriculum development and learning resources should be designed and deployed.

In the next phase of this work and building on current work around creative educational spaces, I propose to investigate the opportunities to re-conceptualize **tangible user interfaces as designed interventions.** Future research in this field might also include longer term qualitative studies to: (1) test **tangible user interfaces** in specific HCI contexts—living laboratories—with particular emphasis on their capacity to scaffold collaboration, creativity, and diversity; (2) further explore their agency in learning ecologies and (3) their impact on the HCI classroom space; (4) investigate specific benefits and challenges for instructors that decide to use the proposed model, coupling multi-pedagogical framework with **tangible user interfaces;** (5) look at how teachers can be scaffolded from a professional development perspective in their deployment of HCI tangibles; and (6) what type of learning resources can be designed for **tangible user interfaces** to deeply enrich the HCI classroom experience.

**ACKNOWLEDGMENTS**

My special thanks go to Prof. Patrick Dillon for sharing with me the adaptive educational environments journey. Thank you also to Mary Featherston for her contagious spaces-for-learning-enthusiasm. Finally, many thanks to Dr. Peter Burrows for our discussions and shared visions, which truly enrich my work.

**REFERENCES**


Tangible User Interfaces as Mediating Tools within Adaptive Educational Environments

cial Interest Group on Computer Graphics, Los Angeles, CA.


*This work was previously published in Enhancing Learning Through Human Computer Interaction, edited by E. McKay, pp. 178-191, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 3.26
Facilitating E-Learning with Social Software:
Attitudes and Usage from the Student’s Point of View

Reinhard Bernsteiner
University for Health Sciences, Medical Informatics and Technology, Austria

Herwig Ostermann
University for Health Sciences, Medical Informatics and Technology, Austria

Roland Staudinger
University for Health Sciences, Medical Informatics and Technology, Austria

ABSTRACT
This article explores how social software tools can offer support for innovative learning methods and instructional design in general, and those related to self-organized learning in an academic context in particular. In the first section, the theoretical basis for the integration of wikis, discussion forums, and Weblogs in the context of learning are discussed. The second part presents the results of an empirical survey conducted by the authors and explores the usage of typical social software tools that support learning from a student’s perspective. The article concludes that social software tools have the potential to be a fitting technology in a teaching and learning environment.

INTRODUCTION
One major task of higher education is to train students for the requirements of their future work by applying and adapting their knowledge to specific workplace-related requirements and settings. Due to the ongoing pressure on enterprises to cut costs, the periods of vocational adjustment in a company will become shorter and shorter.
Facilitating E-Learning with Social Software

On the one hand, the rising pressure of innovation and fast-paced development in the economy results in increased demand for continuous employee training. On the other, growing global competition forces enterprises to use available resources very economically so that employee training is considered to be necessary and desired even though it is conducted under considerable time and cost pressure (Köllinger, 2002).

According to these goals, the settings of the education must be changed adequately: “While most of higher education still ascribes to traditional models of instruction and learning, the workplace is characterized by rapid changes and emergent demands that require individuals to learn and adapt in situ and on the job without the guidance of educational authorities” (Sharma & Fiedler, 2004, p. 543).

In the field of higher education, it has become an important goal to develop “digital literacy” and educate learners as competent users and participants in a knowledge-based society (Kerres, 2007), but it can be assumed that there is a new generation of students, the “digital natives,” who are accustomed to digital and Internet technology (Prensky, 2001a, 2001b).

Oblinger and Oblinger (2005) characterize next-generation students (called “n-gen,” for Net generation) as digitally literate, highly Internet savvy, connected via networked media, used to immediate responses, preferring experiential learning, highly social, preferring to work in teams, craving interactivity in image-rich environments, and having a preference for structure rather than ambiguity.

According to a study conducted by Lenhart and Madden (2005), half of all teens in the USA may be considered “content creators” by using applications that provide easy-to-use templates to create personal Web spaces.

Classical face-to-face learning is seen as rigid and synchronous, and it promotes one-way (teacher-to-student) communication. Thus, it is not surprising that more and more students are opting for Web-based education as a more flexible and asynchronous mode (Aggarwal & Legon, 2006).

The higher education system should provide answers to this new generation of students who enter the system with different backgrounds and skills. They are highly influenced by social networking experiences and are able to create and publish on the Internet (Resnick, 2002).

Educators and teachers therefore have to consider the implications of these developments for the future design of their courses and lectures.

In 2002, a new term, “social software,” entered the stage to refer to a new generation of Internet applications. One focus of this new generation is the collaboration of people in sharing information in new ways such as through social networking sites, wikis, communication tools, and folksonomies (Richter & Koch, 2007).

Wikis, Weblogs, and discussion forums will play a central role in the new context, so the areas of application and possibilities will enlarge enormously. It can be assumed that this will also have considerable influence on learning and the usage of these instruments as learning tools.

This article presents the results of an empirical survey in order to highlight the benefits of the above-mentioned Web-based social software tools from the student’s point of view; 268 first-semester students, all in the first term of their studies at Austrian universities from different study programs, took part in this survey. The students were asked to use one or more of these tools as a learning tool. Participation in this survey was voluntary.

The presentation of the results of this survey is divided into three parts: first, the use of the tools by the students (before they started their studies); second, the experiences the students had made with the tools during the study; and third, the potential future usage.

The article concludes with a discussion of the results of this survey in contrast with other empirical studies already published. Also, the
Facilitating E-Learning with Social Software

limitations of this survey and ideas for further research are pointed out.

THEORETICAL FRAMEWORK

This part refers to the necessary theoretical background required for the following empirical study, especially the areas of social software and learning.

Social Software

The term social software emerged and came into use in 2002 and is generally attributed to Clay Shirky (2003). Shirky, a writer and teacher on the social implications of Internet technology, defines social software simply as “software that supports group interaction.”

Another definition of social software can be found in Coates (2005), who refers to social software as “software that supports, extends, or derives added value from human social behaviour.”

Users are no longer mere readers, audiences, or consumers. They have the ability to become active producers of content. Users can act in user and producer positions and they can rapidly change the position.

Nowadays the term social software is closely related to “Web 2.0.” The term Web 2.0 was introduced by Tim O’Reilly (2005), who suggested the following definition:

Web 2.0 is the network as platform, spanning all connected devices; Web 2.0 applications are those that make the most of the intrinsic advantages of that platform: delivering software as a continually updated service that gets better the more people use it, consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others, creating network effects through an “architecture of participation,” and going beyond the page metaphor of Web 1.0 to deliver rich user experiences.

Web 2.0 technologies such as blogs, wikis, podcasts, and RSS feeds or discussion forums have been dubbed social software because they are perceived as being especially connected and allow users to develop Web content collaboratively and publicly (Alexander, 2006).

Until now, the Internet (Web 1.0) has one big disadvantage: It is easy to get information in it, but it is quite complicated and inconvenient to act as an author and take part in the development of content. Web 2.0 should enable all Internet users to actively take part in the further development of the Internet. Everyone should be able to contribute easily. The focus of Web 2.0 is on the behaviour of the user. It should empower people to communicate, collaborate, contribute, and participate.

This growing phenomenon is very interesting and ought to be examined carefully in order to understand how the Web is evolving and how this continuously regenerative cycle of performance and technological innovation empowers “learning by sharing” (Thijssen & Vernooij, 2002).

Based on the key principle of the architecture of participation, social software can be seen as part of Web 2.0. Wikis, Weblogs, and discussion forums are tools that are seen as social software applications and were selected for further research and the empirical study presented below.

Related Empirical Research

Institutions in the field of higher education have made efforts to introduce various IT-supported learning tools in the daily routine of students and lecturers (Aggarwal & Legon, 2006; Dooley & Wickersham, 2007; Duffy & Bruns, 2006; Evans & Sadler-Smith, 2006; McGill, Nicol, Littlejohn, Grierson, Juster, & Ion, 2005).

Published results of the usage of Weblogs in the Prolearn project (http://www.prolearn-project.org) have shown that a large majority of respon-
students considers personalization and adaptation of the learning environment as important and crucial factors. Learning should be individualized to become more effective and efficient. Personalization is a key element of the learning process, and specific problems need specific solutions as students differ greatly in their backgrounds and capabilities.

Learning materials are typically too general in order to cover a very wide range of purposes and personal learning needs. Compared to classical learning, personalization can be the most important added value that e-learning can offer. With it, education can be optimized and adjusted to various working conditions and needs because students have different goals, interests, motivation levels, learning skills, and endurance (Klamma et. al., 2006).

Chao (2007) explored the potential uses of wikis in the field of software engineering (38 participants), especially for software project team collaboration and communication. Overall, 25 students agreed and 1 student disagreed (2 were neutral) that the wiki is a good tool for project collaboration. Concerning the applications of wikis, more than 23 students found that a wiki is a good tool for maintaining a group diary, managing user stories (project requirements), and project tracking and reporting. While a majority of students found that a wiki is a good tool for updating a project plan, managing acceptance tests, tracking defects, and developing user documents, there were also a significant number of students who disagreed.

First results using wikis for collaborative writing (about 40 participants) also reported similar results. In this study, students used wikis to write articles partly together with the lecturer.

After early problems with using the software and writing contributions in the wiki, students were able to write articles by themselves or in teams. The motivation among students was on different levels, so the lecturer had to increase it during lessons. Other students, however, were highly motivated and were creating the content and adding them to the wikis (Bendel, 2007).

**Constructivism and Learning: Presentation of the Learning Model**

A constructivist point of view focuses on the learning process by looking at the construction of knowledge by an individual. As a consequence, there is a recommendation to align learning environments, especially in the academic context, and associated complex learning objectives with constructivist learning principles (Du & Wagner, 2005; Jonassen, Mayes, & McAleese, 1993; Tangney, FitzGibbon, Savage, Mehan, & Holmes, 2001). Learning is not seen as the transmission of content and knowledge to a passive learner. Constructivism views learning as an active and constructive process that is based on the current understanding of the learner. Learning is embedded in a social context and a certain situation (Schulmeister, 2005).

The constructivist approach shifts learning from instruction and design centered to learner-centered learning and teaching. The role of the educator changes from directing the learner toward supporting and coaching the learner.

Baumgartner (2004) has suggested three different prototypical modes of learning and teaching. These three different modes of learning and teaching can be neutral or specific so they can be applied across all subject domains. Therefore, each teaching model can be used to teach, for example, sociology subjects as well as to teach technical sciences. Learning can be portrayed as an iterative process that can subsequently be subdivided into different phases, which are summarized in Figure 1.

In particular, these three different prototypical modes for learning encompass the following.
Learning and Teaching I: Transferring Knowledge

At the starting point, the learner needs to be provided with abstract knowledge to lay the theoretical foundations and to understand relevant signposts, road markings, and orientation points. This kind of factual knowledge is static and has little value by itself in real and complex situations. It merely serves as a shortcut to prevent pitfalls and to help to organize the student’s learning experiences.

The knowledge of the student is based on knowledge possessed by the teacher. Students have to learn what teachers ask them to learn. The teacher has the responsibility to make the knowledge transfer as easy as possible.

Learning and Teaching II: Acquiring, Compiling, and Gathering Knowledge

In this section of the individual learning career, the student actually applies the abstract knowledge and gathers his or her own experiences. In order to limit the action and reflection possibilities, the learner interacts within a somewhat restricted, artificial environment, which is reduced in complexity and easy to control by the teacher. To provide feedback, the learning environment is designed to include relevant devices where students can deposit their interim products and teachers can inspect them.

The emphasis in this model lies on the learning process of the student. Teachers try to help the students overcome wrong assumptions and wrong learning attitudes, and assist in the reflection process of the subject domain.

Learning and Teaching III: Developing, Inventing, and Constructing Knowledge

Teacher and learner work together to master problems. This model includes problem generation and/or invention. The environment is constructed in such a way that it represents, at least in certain aspects, reality or reality in a constrained form.

Figure 1. Prototypical modes of learning and teaching (Baumgartner, 2004)
Facilitating E-Learning with Social Software

This model includes two-way communication on equal terms, using either linguistic representations or other adequate kinds of language.

Teaching III has strong links to constructivism. From a constructivist point of view, learning is considered as an active process in which people construct their knowledge by relating it to their previous experiences in complex and real situations in life. In their practical lives, people are confronted with unique, unpredictable situations whose inherent problems are not readily observable (Baumgartner, 2004).

Students should be enabled to invent new things, and produce or generate new knowledge. Consequently, learning and teaching at universities in most cases can be assigned to the requirements presented in Learning and Teaching II and III. In order to achieve this goal, a special learning environment must be provided.

Consequences for IT-Supported Learning and Teaching

Computer software can be used for all three models, ranging from programmed instruction (Learning/Teaching I) to problem-solving software (Learning/Teaching II), to complex simulations and/or so-called micro worlds (Learning/Teaching III). It is said that the inherent nature of the Internet brings the real world into the classrooms, and with its hyperlink structure it clearly advocates the model of Teaching III (Baumgartner, 2004).

The use of the Internet, especially through its social software, gains importance because it can contribute to exceed the limits of classical teaching models. By adapting learning and teaching models to the new technical possibilities, the roles of learner and teacher are becoming more indistinct because the learner can take a central part in the design and arrangement of the learning process (Kerres, 2006).

Systems that support learners with respect to the Learning Model III are called personal learning environments (PLEs). PLEs are mostly Web-based applications and are based on learning management systems (LMSs; Seufert, 2007).

PLEs are personal and open learning environments, and they are suitable for cross-linking content and people. Learners can use PLEs to manage individual learning progress. They are ideally available for lifelong learning and are supported by the following processes.

• setting up individual learning goals
• planning and controlling one’s own learning concerning the content as well as the learning process
• combining formal and informal learning activities
• communicating with peers during the learning process
• establishing social networks or communities of practice
• using Web-based services, for example, syndication
• verifying the learning process with respect to the learning goals

Unlike an LMS, which is usually related to one special institution or to one special course, a PLE is focused on the individual learner. A PLE should combine a broad mixture of different resources and subsystems in a “personally-managed space” (Attwell, 2006).

In the previous decade, learning management systems were developed that moved toward enterprise-level applications, “but the wealth of new, user-friendly, tools in the Web 2.0 environment suggests that the all-in-one monolithic e-learning systems may be entering a phase of obsolescence by the ongoing development of the web” (Craig, 2007).

Social software applications have the potential to cope with these requirements (Brahm, 2007).
DESCRIPTION AND CLASSIFICATION OF SOCIAL SOFTWARE TOOLS

In the following section, three social software tools—Weblogs, discussion forums, and wikis—are described in more detail and the tools are compared. Students were able to select these tools during the empirical study.

Weblog

A Weblog, a compound of Web and logbook, usually just called “blog,” is a Web site that contains new articles or contributions in a primarily chronological order, listing the latest entry on top.

Primarily, a Weblog is a discussion-oriented instrument especially emphasizing two functions: RSS feeds and trackback. RSS feeds, also called RSS files, can be read and processed for further use by other programs. The most common programs are RSS readers or RSS aggregators that check RSS-enabled Web sites on behalf of the user to read or display any updated contribution that can be found. The user can subscribe to several RSS feeds. Thus, the information of different Web sites can be retrieved and combined. Preferably, news or other Weblogs are subscribed to.

Trackback is a service function that notifies the author of an entry in a Weblog if a reference to this contribution has been made in another Weblog. By this mechanism, a blogger (person who writes contributions in a Weblog) is immediately informed of any reactions to his or her contribution on other Weblogs (Hammond, Hannay, & Lund, 2004).

Forum

A discussion forum or Web forum is a service function providing discussion possibilities on the Internet. Usually, Web forums are designed for the discussion of special topics. The forum is furthermore subdivided into subforums or subtopics. Contributions to the discussion can be made and other people may read and/or respond to them. Several contributions to a single topic are called a thread.

The application areas of the two instruments, Weblogs and forums, are quite similar. The most essential differences between Weblogs and discussion forums can be described as follows:

• A forum is usually located on one platform while many bloggers develop their own, individual environments. They connect their Weblogs via RSS feeds and trackback functions.
• Through the integration of RSS files and trackback functions, a discussion process can be initiated and continued, crossing the boundaries of the bloggers’ own Weblogs without authors having to observe other Weblogs.
• Weblogs tend to be more people centered whereas forums are more topic focused. Through the use of Weblogs, learner-specific learning environments can be constructed without interfering with the learning environments of others (Baumgartner, 2004).

Wiki

A WikiWikiWeb, shortly called wiki, is a hypertext system for storing and processing information. Every single site of this collection of linked Web pages can be viewed through a Web browser. Furthermore, every site can also be edited by any person. The separation between authors and readers who write their own text, and change and delete it is obsolete as also third parties can carry out these functions (Augar, Raitman, & Zhou, 2004).
Facilitating E-Learning with Social Software

Learning Activities Supported by Social Software

The integration of different social software tools offers support in the following learning activities:

- **Learning from different perspectives**: The integration supports the exchange of ideas as well as finding like-minded people. Furthermore, social software tools simplify the process of establishing connections between people of the same or similar interests. Simultaneously, its open and expandable philosophy supports going beyond the thinking in groups (of a common interest) by supporting diversity and bringing together different perspectives and backgrounds (Efimova & Fiedler, 2004; Schulmeister, 2004).

- **Synergies of self-organized and joint learning**: Social Software tools provide a personal learning area for their authors. However, this does not force a general learning flow or learning style. Learners are not alone and can profit from the feedback of a community in order to examine and enhance the development of their own ideas (Böttger & Röll, 2004; Efimova & Fiedler, 2004; Fiedler, 2004).

- **Digital apprenticeship**: Through reading other wikis, forums, or Weblogs regularly, beginners are enabled to learn from experts. At the same time, they can actively participate in discussions beyond geographic or thematic borders (Efimova & Fiedler, 2004; Fiedler, 2004).

- **Support for the development of the ability to learn (learning to learn)**: Through the publication of one’s own thoughts and reflections, content is made available for assessment as well as for further development, thereby improving self-observation and self-reflection skills. The knowledge change of the learner will be improved (Baumgartner, 2005).

- **Support for reflexive writing**: The simple but efficient and rather robust encoding standard usually used in social software allows for the explicit modeling of content flows, feedback loops, and monitoring procedures of various kinds, thus supporting an ongoing iterative process of explication and reflection (Fiedler, 2004).

Integration of Social Software Tools and the Learning and Teaching Modes

Baumgartner (2004) has integrated different types of content management systems in relation to the most suitable learning and teaching mode. He clearly states that the boundaries are overlapping and that every tool, in one way or another, could be used for every teaching model. Figure 2 presents the integration of the social software tools and the learning and teaching modes.

Weblogs and forums can be defined as discussion-oriented tools because the discourse and exchange of ideas related to a certain topic is the preeminent aim. Weblogs offer the possibility to support all three phases of the learning process. However, the main focus can be assigned to Modes II and III.

Based on the multitude of interaction possibilities, wikis can be attached to Teaching III (Baumgartner, 2004). Additional functions were added to Weblog tools that go beyond the scope of the central use of Weblogs; for example, longer articles can also be stored. Through the creation of directories, a structured collection of links can be implemented.

Through the additional linking of Weblogs, wikis, and forums, there is the possibility to develop a personal knowledge collection (Kantel, 2003).
Facilitating E-Learning with Social Software

The purpose of this survey was to determine if the integration of Web-based social software tools (wikis, discussion forums, and Weblogs) are suitable to foster learning from the student’s point of view.

Aim of the Survey and Methodology

Scrubbing the possibilities and constraints of social software tools (wikis, discussion forums, and Weblogs) as personal learning environments, students at Austrian universities were asked to use one or more of the offered tools for their research, homework, and documentation purposes. In most cases, the collaboration of students was required to perform the assigned tasks.

The students were asked to use the tools for one course only during the winter term of 2006. Furthermore, there was no obligation for the students to use a tool at all; they were just encouraged to do so. Students were also offered the possibility to use two or three tools; the selection was up to the students.

The courses were organized as blended-learning courses so they included on-campus lessons and off-campus work in which the students could work face-to-face or using the social software tools.

More than 90% of the students attending the courses took part in this survey. In order to give the participants an impression of the functionality and usage of the tools, short presentations of the tools were made by an instructor before the students made their choice.

At the end of the testing phase—after 4 weeks of using the tools—selected students reported their experiences with the tools used. Students who had decided not to use the tools in the first place got an impression about the usage, advantages, and disadvantages of the tools from their fellow students. Following these short presentations, a questionnaire was completed that provided the basic findings for further inspection and research.

A total of 268 first-semester students of different Austrian universities in five selected courses took part in this survey. The majority of the participants were between 18 and 20 years old. The portion of female students was about 17%.

Figure 2. Prototypical modes and social software tool
According to a survey conducted by Seybert (2007) concerning gender differences in computer and Internet usage by young people (aged between 16 and 24), there is no gap between men and women in Austria. The proportion of women and men (in the relevant age class) that use a computer (almost) once a day is 72% the same. A study by Otto, Kutscher, Klein, and Iske (2005) indicates that there is a positive correlation between a formal educational background and the usage of the Internet in Germany: “Beside socio-cultural resources like family background, peer structures and social support in general, the formal educational background turns out to be the main factor for explaining differences in internet usage” (p. 219). As a consequence, for the analysis of the results of this survey, no distinction between male and female students was made.

Table 1 presents the distribution of the participants concerning the degree programs the students are attending.

For the further analysis of the results, no distinction according to degree programs will be made.

The questionnaire asked each participant questions about her or his subjective impression of the application of the tools. It included 5-point Likert scales for rating constructs such as eligibility, perceived quality, or enjoyment.

The study was conducted to find answers about the:

- usage of social software before the study started,
- selection of the offered tools,
- perceived quality of the contributions and the support for learning,
- applicability of the instruments to support communication and community building,
- correlation of the usage for private and educational purposes of the tools,
- fun factor in using the instruments, and
- potential future usage.

The results of the study are presented in three parts:

- Part 1: Analysis of the usage of wikis, discussion forums, and Weblogs of the students before the study was started
- Part 2: Experiences made with the tools during the study
- Part 3: Potential future usage of the tools

### Part 1: Tool Selection and Prestudy Usage

Due to the fact that the students could select the tools on their own, the Table 2 shows the results of this selection process.

According to Table 2, the combination of wikis and discussion forums is the most selected combination of tools (42.9%), followed by wikis only (23.1%) and discussion forums only (22.4%). In the end, only five students (1.9%) did not take part in the study; they did not select a tool, although they first had the intention to do so. Only one student used Weblogs only. Generally, Weblogs were not used very intensively by the participants.

Table 3 shows the usage of the tools by the participants before they took part in the study. It indicates that wikis (76%) and discussion forums (78%) are currently the most widely used.
Facilitating E-Learning with Social Software

Table 2. Tools selected by the students

<table>
<thead>
<tr>
<th>Only one tool selected</th>
<th>Percent</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikis only</td>
<td>23.1%</td>
<td>62</td>
</tr>
<tr>
<td>Discussion forums only</td>
<td>22.4%</td>
<td>60</td>
</tr>
<tr>
<td>Weblogs only</td>
<td>0.4%</td>
<td>1</td>
</tr>
<tr>
<td>More than one tool selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wikis and discussion forums</td>
<td>42.9%</td>
<td>115</td>
</tr>
<tr>
<td>Wikis and Weblogs</td>
<td>1.9%</td>
<td>5</td>
</tr>
<tr>
<td>Discussion forums and Weblogs</td>
<td>0.7%</td>
<td>2</td>
</tr>
<tr>
<td>Wikis, discussion forums, and Weblogs</td>
<td>6.7%</td>
<td>18</td>
</tr>
<tr>
<td>No tool selected</td>
<td>1.9%</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3. Students already using the tools

<table>
<thead>
<tr>
<th>Wiki</th>
<th>Forum</th>
<th>Weblog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>76%</td>
<td>78%</td>
</tr>
<tr>
<td>No</td>
<td>24%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Table 4. Usage for private purposes

<table>
<thead>
<tr>
<th>Statement</th>
<th>Wiki</th>
<th>Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>35%</td>
<td>29%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>9%</td>
<td>9%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>16%</td>
<td>17%</td>
</tr>
<tr>
<td>I disagree</td>
<td>8%</td>
<td>12%</td>
</tr>
</tbody>
</table>

Table 5. Usage for educational purposes

<table>
<thead>
<tr>
<th>Statement</th>
<th>Wiki</th>
<th>Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>57%</td>
<td>22%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>33%</td>
<td>29%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>3%</td>
<td>12%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>8%</td>
<td>24%</td>
</tr>
<tr>
<td>I disagree</td>
<td>1%</td>
<td>12%</td>
</tr>
</tbody>
</table>
Table 5 presents the results for the statement “I often use wikis or forums for educational purposes.”

A huge majority (90%) stated that they use wikis for educational purposes and about two thirds (68%) used wikis for private purposes. Wikis are therefore more intensively used for educational purposes than for private purposes, whereas the usage of forums is exactly the opposite: They are more used for private purposes than for education.

The responses of the students concerning these statements were that wikis are foremost considered as a source of serious information, whereas forums are ideal for getting hints or clues to problems because of their privacy. Questions about computer problems, computer games, leisure activities, and so forth were mentioned. A repetition of this image can be identified when the disagreement with the statement is analyzed; 29% of the students do not or rarely use forums for private purposes compared to 36% for education.

**Part 2: Experiences Made During the Study**

This section presents the results of the study concerning experiences with the usage of the tools during the study.

### Quality and Support for Learning

The following refers to statements concerning the quality of the contributions of wikis and discussion forums and their support for learning.

The results of the statement “The quality of contributions in wikis or forums is in general good” are presented in Table 6. The contributions of wikis are evaluated to be much better than those of forums.

The surveyed pupils had the opportunity to give reasons for their assessment concerning the quality of contributions via additional qualitative feedback. The following summarizes the addressed reasons.

One reason for the excelling grade for the quality of wikis is the “Wikipedian community.” The term *wiki* is often seen as a synonym for the free online encyclopedia Wikipedia (http://www.wikipedia.org). Wikipedia is widely used for a great variety of tasks, including research on all topics needed for educational and private purposes.

In contrast to the good evaluation of the contributions of wikis, the open architecture of wikis was also mentioned. In most cases, this open architecture allows everyone to edit entries, which results in the uncertainty of whether the knowledge presented is correct or not. The quality of contributions in discussion forums was rated

<table>
<thead>
<tr>
<th></th>
<th>Wiki</th>
<th>Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>38%</td>
<td>10%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>52%</td>
<td>31%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>10%</td>
<td>41%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>2%</td>
<td>15%</td>
</tr>
<tr>
<td>I disagree</td>
<td>0%</td>
<td>4%</td>
</tr>
</tbody>
</table>

### Table 7. Clarity of contributions

<table>
<thead>
<tr>
<th></th>
<th>Wiki</th>
<th>Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>2%</td>
<td>4%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>6%</td>
<td>18%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>29%</td>
<td>37%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>34%</td>
<td>27%</td>
</tr>
<tr>
<td>I disagree</td>
<td>29%</td>
<td>14%</td>
</tr>
</tbody>
</table>
Facilitating E-Learning with Social Software

rather mediocre. Forums are primarily used for technical problems, especially computer-related problems; to get in contact with experts on certain topics, and to get information on online games.

The next statement, “The usage of wikis or forums leads to misunderstandings and confusion,” is about the clarity of the contributions.

Only a minority think that the contributions are not clear and may lead to misunderstandings. In this case, wikis are also rated better than forums.

The next statements addressed the support of these instruments for learning. Table 8 summarizes the results for the statement “When reading contributions in wikis or forums, it is easier for me to acquire the learning contents.”

More than half of the students express that reading contributions in wikis is helpful for learning, whereas only about 8% think that it is not helpful. Compared to forums, wikis were again much better evaluated, especially considering the big difference from the negative evaluations of forums.

Table 9 presents the learning support achieved by writing contributions. (“When writing contributions in wikis or forums, it is easier for me to acquire the learning contents.”)

A different picture emerges in the statistics when comparing the evaluation of how writing an article or post supports the learning process. Here, forums take the lead when it comes to positive assessment. In both cases, there was a large number stating that writing is neither positive nor negative. The majority of the students read rather than wrote, while more students wrote in forums than in wikis.

Table 8. Reading contributions helps to acquire contents

<table>
<thead>
<tr>
<th></th>
<th>Wiki</th>
<th>Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>23%</td>
<td>8%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>36%</td>
<td>21%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>32%</td>
<td>31%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>5%</td>
<td>25%</td>
</tr>
<tr>
<td>I disagree</td>
<td>3%</td>
<td>15%</td>
</tr>
</tbody>
</table>

Table 10. Applicability for communication

<table>
<thead>
<tr>
<th></th>
<th>Wiki</th>
<th>Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>9%</td>
<td>39%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>33%</td>
<td>37%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>29%</td>
<td>17%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>15%</td>
<td>4%</td>
</tr>
<tr>
<td>I disagree</td>
<td>15%</td>
<td>3%</td>
</tr>
</tbody>
</table>

Table 9. Writing contributions helps to acquire contents

<table>
<thead>
<tr>
<th></th>
<th>Wiki</th>
<th>Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>8%</td>
<td>7%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>13%</td>
<td>19%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>45%</td>
<td>34%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>14%</td>
<td>22%</td>
</tr>
<tr>
<td>I disagree</td>
<td>19%</td>
<td>17%</td>
</tr>
</tbody>
</table>

Table 11. Support for community building

<table>
<thead>
<tr>
<th></th>
<th>Wiki</th>
<th>Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>10%</td>
<td>28%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>25%</td>
<td>32%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>39%</td>
<td>23%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>15%</td>
<td>11%</td>
</tr>
<tr>
<td>I disagree</td>
<td>10%</td>
<td>6%</td>
</tr>
</tbody>
</table>
Facilitating E-Learning with Social Software

Applicability for Communication and Community Building

The statement was formulated as follows: “Wikis or forums are appropriate to support communication.”

The results clearly demonstrate that discussion forums are made for communication whereas wikis are rather seen as a kind of reference book or encyclopedia, as already mentioned above.

The results of the next statement, “Wikis or forums support the setup of communities,” can be seen in Table 11.

Opinions about the applicability of wikis to establish a community is split. About 35% say that wikis are supportive of building a community compared to 25% who said that wikis do not support community building. The support of forums to build a community is rated much better: 50% indicated that forums are well suited to build a community. These results were to be expected because they confirm the nature of the instruments.

Fun Factor in Using the Instruments

In surveying whether students gain pleasure (“I enjoy using wikis or forums”), wikis again came out on top.

A majority enjoy using wikis (62%) and forums (56%). Considering the percentage of students who said that there is no (“I disagree”) or little (“I slightly disagree”) fun when using these instruments, wikis (6%) are much better rated than forums (18%).

Part 3: Potential Future Usage of the Tools

The third section of the empirical study deals with the potential usage by students who had not used the instruments before the study. Students gained knowledge and experiences by using the tools during the study by themselves or on the basis of the reported experiences made by their fellow students.

Table 13. Future usage in educational context (current nonusers)

<table>
<thead>
<tr>
<th></th>
<th>Wikis</th>
<th>Forums</th>
<th>Weblogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>18%</td>
<td>16%</td>
<td>13%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>36%</td>
<td>23%</td>
<td>23%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>30%</td>
<td>16%</td>
<td>24%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>9%</td>
<td>12%</td>
<td>13%</td>
</tr>
<tr>
<td>I disagree</td>
<td>7%</td>
<td>33%</td>
<td>27%</td>
</tr>
</tbody>
</table>

Table 14. Future usage in private context (current nonusers)

<table>
<thead>
<tr>
<th></th>
<th>Wikis</th>
<th>Forums</th>
<th>Weblogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>11%</td>
<td>14%</td>
<td>9%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>36%</td>
<td>23%</td>
<td>22%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>30%</td>
<td>25%</td>
<td>24%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>14%</td>
<td>7%</td>
<td>16%</td>
</tr>
<tr>
<td>I disagree</td>
<td>9%</td>
<td>32%</td>
<td>28%</td>
</tr>
</tbody>
</table>

Table 12. Fun Factor in using the instruments

<table>
<thead>
<tr>
<th></th>
<th>Wiki</th>
<th>Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>I totally agree</td>
<td>26%</td>
<td>19%</td>
</tr>
<tr>
<td>I generally agree</td>
<td>36%</td>
<td>37%</td>
</tr>
<tr>
<td>Neither...nor (neutral)</td>
<td>31%</td>
<td>26%</td>
</tr>
<tr>
<td>I slightly disagree</td>
<td>5%</td>
<td>14%</td>
</tr>
<tr>
<td>I disagree</td>
<td>1%</td>
<td>4%</td>
</tr>
</tbody>
</table>
The first statement, “I will use wikis, forums, or Weblogs for educational purposes in the future,” yielded the results in Table 13.

According to this study, wikis will have a bright future and will be used often for educational purposes, whereas forums will be used less often. About 54% of the surveyed students had the intention of using wikis more or less often in the future. About 16% did not think that they will use wikis often in the future and 30% are not yet sure if they will use this instrument.

The results for forums and Weblogs indicate no clear trend, but forums were rated slightly higher than Weblogs; 39% of the students stated that they can imagine using forums in the future for their education compared to 36% for Weblogs. At the other end of the scale, 45% did not have the intention to use forums (40% for Weblogs).

The statement “I will use wikis, forums, or Weblogs for private purposes in the future” leads to similar results.

From this point of view, wikis are again the leading instrument, followed by forums and then Weblogs.

It must be said that the responses to this set of statements represented feelings, attitudes, and opinions about instruments that had not yet been used by the asked participants. The purpose of posing these statements was to gain insight into the mind-set in regard to these instruments.

**DISCUSSION**

The results clearly show that wikis are currently the most often used instrument and furthermore have the greatest potential as a tool for learning and knowledge management in the field of learning; these findings are in line with other empirical studies (Bendel, 2007; Chao, 2007).

Other studies (McGill et al., 2005; Nicol & MacLeod, 2004) report that a shared workspace helps to support collaborative learning; the possibility of being able to access and contribute to the development of resources at any time and from any location was especially appreciated by the students.

The survey at hand made a distinction between reading and writing contributions to wikis and discussion forums. The results show that 59% of students said reading contributions in wikis is helpful for learning (stating “I totally agree” and “I generally agree”) while only 21% stated that writing in wikis is helpful for learning. Reading contributions in forums helped 29% of the participants, whereas writing in forums was helpful to 26%. This survey supports the general statement that a shared workspace that supports a constructivist and learner-centered approach is helpful for learning.

The pedagogical value in the context of learning is described in several publications (Babcock, 2007; Hurst, 2005). Weblogs can foster the establishment of a learning and teaching environment in which students and teachers experience a greater degree of equality and engagement. Du and Wagner (2007) published results of a study of an information systems undergraduate course (31 participants). This study indicated that the performance of students’ Weblogs was a significant predictor for learning outcomes, while traditional coursework was not. Moreover, individuals’ cognitive construction efforts to build their own mental models and social construction efforts to further enrich and expand knowledge resources appeared to be two key aspects of constructivist learning with Weblogs. According to this study, there is a potential benefit of using Weblogs as a knowledge construction tool and a social learning medium.

According to the survey at hand, Weblogs are not yet widely used, and their potential seems to be limited. It can be assumed that these limited prospects will change when the penetration of Weblogs into the daily routine of the students increase—for private as well as for educational purposes.
To avoid possible pitfalls about the application of these instruments in the context of learning, some social and psychological issues must be taken into consideration (Kreijns, Kirschner, & Jochems, 2003). Social interaction is essential for members of a team to get to know each other, commit to social relationships, develop trust, and develop a sense of belonging in developing a learning community. The size and the composition of the learning communities seem to be important factors in how interaction and communication within the learning community will take place (Dooley & Wickersham, 2007). There are also many unresolved issues, like the provision of the technology and the services, intellectual property rights and digital rights management, the security of data, access restrictions to the content, and information ethics (Attwell, 2006; McGill et al., 2005; Sharma & Maleyeff, 2003).

CONCLUSION

The aim of this contribution was to investigate the experiences of students using social software tools in the context of learning. Wikis, Weblogs, and discussion forums are typical social software tools and were used for this survey.

The results clearly show that wikis and discussion forums can support learning and collaboration. The usage of Weblogs in this study was limited and hence no statements about their applicability can be made. In order to assure a successful implementation of these tools, social and psychological issues must be taken into consideration as well.

The results of this study are the basis for the introduction of social software into education to help students set up individual learning environments. These learning environments should support lifelong learning.

There are likely to be other unplanned consequences of the intensive use of the Internet in general and social software especially. Further research is needed to explore possible problems and solutions.

The results of the empirical survey indicate that a long-term study in combination with the further development of social software tools may be promising.

REFERENCES


Facilitating E-Learning with Social Software

(Ed.), ePortfolio Forum Austria: Proceedings of the ePortfolio Austria 2005, Salzburg, Austria (pp. 33-38).


Facilitating E-Learning with Social Software

http://www.dlib.org/dlib/december04/hammond/12hammond.html


Facilitating E-Learning with Social Software


Thijssen, T., & Vernooij, F. (2002). Breaking the boundaries between academic degrees and lifelong learning. In T. A. Johannessen (Ed.),
Facilitating E-Learning with Social Software


This work was previously published in the International Journal of Web-Based Learning and Teaching Technologies, edited by L. Esnault, Volume 3, Issue 3, pp. 16-33, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Section IV
Utilization and Application

This section introduces and discusses the ways in which particular technologies have impacted humanity and proposes new ways in which IT-related innovations can be implemented within organizations and in society as a whole. These particular selections highlight, among other topics, ubiquitous computing applications in an educational setting and computer-mediated communication. Contributions included in this section provide excellent coverage of today’s world and insight into how our interaction with technology impacts the fabric of our present-day global village.
Chapter 4.1
Human Computer Interaction and the Best Mix of Face-to-Face and E-Interactions in Educational Settings

Bolanle A. Olaniran
Texas Tech University, USA

ABSTRACT

Recent trends and rapid improvement in technology such as computer-mediated communication (CMC) and increasing bandwidth in the Internet are facilitating increased electronic interactions (i.e., e-interactions otherwise known as or commonly referred to as the human computer interaction (HCI)). CMC technology systems are a common occurrence in educational institutions as administrators attempt to encourage technology usage and instructors race to learn and implement CMC use in their classrooms and students demand greater flexibility and control in how they learn. Notwithstanding is the need to decide which forms of HCI technology to use, how to use them, and what benefits can accrue from such usage. The discussion here explores each of these issues, but more specifically will focus on addressing the case for blending e-interactions with the traditional face-to-face (FTF) communication medium while addressing the appropriateness of such combination.

INTRODUCTION

Human computer interaction (HCI) occurs through a host of information communication technologies (ICT), specifically computer-mediated systems (e.g., e-mail, computer conferencing, video-conferencing) that facilitate electronic interaction among people. A multitude of organizations involved in the knowledge management industry are finding ways to incorporate computer-mediated communication (CMC) technologies into their day-to-day operations as coordination tools and learning curriculum dissemination tools. HCI has also found explosive growth in educational settings such as traditional and non
traditional universities to extend current instructional content delivery and to tap into distance educational settings. The idea of CMC systems for the instructional delivery tool to provide both uniform and customized training is called electronic learning, which is otherwise referred to as e-learning. Contemporary research has discussed at length the potential benefits of CMC systems in general and within distance education in particular (Barnard, 1997; McIsaac & Gunawardena, 1996; Yakimovicz & Murphy, 1995), but research on joint use of the CMC system and face-to-face (FTF) communication in regular classrooms is scarce (Olaniran, 2004). This chapter will address how using a combined CMC and (FTF) interaction could benefit courses in communication and other social sciences.

Teachers are challenged to incorporate information communication technology (ICT) into their curriculum to enhance learning and to prepare students for future careers (Althaus, 1997; Craig, 2001-2002; Snoeyink & Ertmer, 2001-2002; Witmer, 1998). ICT offers opportunities for facilitating discussion and presenting forums for varieties of opinions (McComb, 1994; Olaniran, 2001; Olaniran, Savage, & Sorenson, 1996, Witmer, 1998). HCI technology, specifically, CMC, offers the opportunity for active and experiential learning and its benefits in group activities have been acknowledged (Craig, 2001-2002; Gasker & Cascio, 2001; Olaniran, 1994; Olaniran, et al., 1996).

In spite of the identified advantages, incorporating ICT especially CMC into classrooms remains challenging. Some challenges facing instructors regarding implementation of CMC in classrooms and courses include selection and usage long after adoption decisions are made. In addition to technology issues, instructors need to focus on pedagogical issues surrounding course structure, and course management.

**COMPUTER MEDIATED COMMUNICATION AND LEARNING FACILITATION**

The rampant effect of information communication technology (ICT) is constantly being felt in contemporary organizations, interpersonal interactions, and academic settings. ICTs are instrumental in facilitating human computer interaction, which underlies the computer mediated communication (CMC) in organizations, classrooms, groups, and interpersonal contexts. As a matter of fact, the issue facing most institutions today (academic and non academic organizations) is not whether to use CMC systems but rather how to use them effectively.

There is a significant volume of literature on CMC systems and distance education (e.g., Abrahamson, 1998; Twigg, 1997, Wegrif, 1998), however, not all usages of CMC in learning environment are exclusively distance education. Some research in CMC technologies concentrates on the social effects of communication technologies (e.g., Hiltz & Turoff, 1978; Sproull & Kiesler, 1986). No doubt, the demand for distance education contributes to the proliferation of ICT as a way to provide education and training. However, there is a need to look at how CMC and other technologies facilitate learning and to identify other key variables that must be accounted for in order for effective learning to occur. This chapter seeks to provide a practical guide to educators and learners for enhancing learning through communication technology.

The globalization trends in contemporary organizations are putting priority on uniform and customized training to the extent that institutions are looking at e-learning to meet their curriculum needs. As a result, varieties of online universities like the University of Phoenix and Westwood College Online are developing in order to meet the needs of non traditional students (e.g., corporate travelers and expatriates). Furthermore, several universities are throwing their support behind
the development of technological centers such as the teaching, learning, and technology center (TLTC) to assist instructors in incorporating CMC into teaching and learning. Specifically, the developments in the use and adoption of ICTs in the educational arena necessitate a closer look at learning that takes place where these technologies are being deployed.

There are several areas to examine when dealing with e-learning or incorporating CMC into courses. These include areas such as: the interaction between students and technology, the interaction between students and teacher, the interaction between students and students. Robson (2000) argues that learning theories are mediated by technologies given that students often establish relationships with varieties of technologies. In essence, it is important to examine how students learn when using CMC systems (Alavi, 1994; Du & Wagner, 2005). Furthermore, it is hoped that examining how students learn and interact with technology will help instructors and, perhaps, designers to determine the best way to apply technology in course curricula.

**HOW STUDENTS LEARN WITH COMPUTER MEDIATED COMMUNICATION SYSTEM**

Perhaps the most prominent reason administrators and instructors offer for incorporating HCI and other CMC technologies into course design is the vision of personalizing learning to meet the idiosyncrasies of students’ learning styles. Innovative teachers see computers as tools to help students in their learning needs with the goal of tailoring course content to suit each student. Similarly, personalized learning is closely related to active or transformational learning in the sense that it allows for students to apply instructional content while retaining information longer than the lecture and print mode that is characterized by traditional modes of instruction (Olaniran, Stalcup, & Jensen, 2000). Students apply computers to learning in several ways (1) to access information, (2) to develop skills, and (3) to cultivate appropriate behaviors and other social skills (Robson, 2000). Each of the uses is explored in the following paragraphs.

**INFORMATION ACCESS**

CMC offers students the opportunity to scan information through Website navigation (browsing) and by searching for particular information. For instance, in Web-assisted methods of CMC, students are able to navigate instructors’ Web pages to access course notes and other instructional resources provided by instructors. However, access to information is not restricted to just instructors’ Web pages. Students can use a browser to seek additional information or to research similar topics and other topics of interest. General browsing is categorized as low cognitive activity in learning (Jacobson, 1995). Browsing can, however, facilitate a high cognitive thinking because students learn as they seek varieties of information and when they develop a method to categorize information in a structured manner (i.e., retention and recall) at a later time (Jacobson, 1995; Robson, 2000). At the same time, one must not forget that students differ even in the way they search and learn. For instance, Robson (2000) argues that some students prefer general browsing to specific keyword searches. Students also differ in the way they avoid disorientation or get lost in the multitude of information (Hara, Bonk, & Angeli, 2000; Wegrif, 1998). For example, some students are able to navigate through multiple sites and keep track of information obtained, while others are stressed and are unable to distinguish one network from another (Duchier, 1996).

As far as information access and personalized learning is concerned, Olaniran (2004) argues that a mixed or combined CMC and face-to-face communication media (CMC/FTF) offers a good
way of addressing students’ needs and at the students’ convenience and comfort levels. In a study evaluating the combined CMC/FTF media, Olaniran (2004) finds that students using CMC were three times more likely to initiate contact with the instructor about course content than in an FTF setting. Specifically, the results also indicate that students are twice as likely in the use of the CMC medium when compared to FTF medium to address questions relating to individualized learning. Thus, the CMC/FTF design structure offers a way to compare both media while presenting a method for accounting for individual students’ interest, knowledge level, and learning styles in a gradual manner (Gasker & Cascio, 2001; Olaniran, et al., 1996). CMC structure allows teachers to target different learning styles and knowledge levels without confusing other students through individualized contacts and course designs.

Other factors influencing access include the propensity to adopt and use communication technologies. Adopting CMC for course design and students’ learning starts with access to technology and the willingness to use it (Olaniran, 1993). A study on attitude and perceptions of e-learning found ease of access as the top reason in students’ use of CMC at 47 percent, while course relevance to future careers and user friendliness were at a distant second and third (i.e., 29 percent and 24 percent respectively) (Vaughan & MacVicar, 2004). This finding reinforces the need for instructors to pay specific attention to issues of accessibility. It is often taken for granted that students have access to e-mail and Internet especially in the USA. This may be obvious because some students have access to internet while on campus whereas they do not have access at home. More troubling on most university campuses is the fact that students often shy away from using a university assigned e-mail account for other third party client e-mail such as Yahoo, MSN, and Hotmail and the likes. The challenge is that some university are clamping down on the use of these third party e-mail accounts as a preventive measure against the spread of viruses and worms to their systems (Curran, McIntyre, Meenan, McCloy, & Heaney, 2004). Consequently, some students are unable to access materials sent by instructors.

**SKILL DEVELOPMENT**

Skills take two forms. The first deals with the technical skills that allow both student and instructor to use hardware and software or evaluate data. The second involves the social skills that are deployed by technology users during interactions (Robson, 2000). The technical skills address the issue of competence with the technology in use, while social skills deal with issues of etiquette and discernment of contextually appropriate behaviors. In CMC research, Olaniran (2004) addresses the adjustment process that accompanies students’ interaction, which range from use of all caps to protocols of posting messages in a threaded discussion. Personal experience in threaded discussions also suggests that students often participate in threaded discussion by breaking the thread; that is, starting a new thread while still on the same topic.

**NURTURING APPROPRIATE ATTITUDES**

As students use CMC they develop attitudes toward the technology medium (Olaniran, 2004) as well as developing professional attitudes and values consistent with the course content or discipline. Robson (2000) argues that the learning of concepts, skills, and development of attitudes equates “cognitive development” on the part of the student, which represents an important goal in all forms of education. However, it is not certain that attitude development toward technology translates positively to professional attitude development. For instance, how users (i.e., students) perceive a
CMC medium based on ease of use affects decisions to adopt and continue using the technology (e.g., Olaniran, 1993; 2004; Vaughan & MaeVicar, 2004). Thus, one could argue that when students are frustrated within a CMC system, they may abandon the technology or fail to adopt it.

At the same time, students never get to develop professional attitudes or achieve high cognitive development when they fail to use a particular communication technology and thus, they rob themselves the opportunity to learn the technology even when they know it might be beneficial to them. For instance, in a course utilizing Web authoring tools, I have found few students deciding to skip certain assignments (designing personal Web page) simply because of the perceived difficulty with the technology software. Students discuss the lack of competence in the Web authoring tools offered through the university computing services center. Therefore, it is imperative for instructors to realize that simply because the tool is available and the instructor has offered some training on the tools does not mean that students grasp the concept and are ready to perform certain tasks on their own. This is especially the case with non-technically oriented courses in liberal arts and social sciences. Robson (2000) echoes this assessment when arguing that “access to knowledge through technology does not guarantee improved outcomes” (p. 157). Instructors are then critical about facilitating students learning; how they design the learning experience is what will facilitate the desired outcomes. This raises the issue of how instructors can assist students in learning, which is discussed in the next section.

CRITICAL THINKING AND TRANSFORMATIVE LEARNING

The use of CMC in teaching courses can foster transformational learning. The transformation learning provides an opportunity to contrast experiential knowledge (i.e., previously acquired information) with new information in order to make informed decisions about which to accept (Olaniran, 2004). The transformational learning identifies three factors (1) Validation of prior knowledge, (2) modification of prior knowledge, and (3) replacement of prior knowledge (Craig, 2001-2002; Salaberry, 2000). The process is facilitated through the direct communication process enhanced in CMC (Alavi, 1994; Mason & Bascich, 1998). Olaniran (2004) reported evidence of transformation learning when students challenged their existing beliefs about both CMC and FTF media, that is, some changed their prior beliefs after they were confronted with new information and experience. The best evidence of transformation learning reported by the learners is illustrated in the understanding of the reason for the change however.

It is possible that the change in opinion of both communication media could be attributed to the course design. The CMC and FTF format helped fortify the transformational learning experience by providing students with practical use of communication technology, which provides students with the opportunity to validate information (beliefs and facts) about both study and the communication media (CMC vs. FTF). The learning

HELPING STUDENTS TO LEARN WITH CMC

Incorporating CMC to a course design is not a panacea. However, how technologies are applied will be the major determinant of the learning that takes place. CMC platforms can be used simply as a conduit to create printed materials or it can be used as an all-encompassing learning forum (access, knowledge, interaction forum etc.). Electronic media provide a platform for interaction and can be a source of knowledge (Du & Wagner, 2005; Leidner & Jarvenpaa, 1995). For instance, the Internet is viewed as a gateway to information (Uys, 1997) and a source of knowledge.
process that occurs can then enhance greater retention of information about course content and attitudes toward technology. Furthermore, interaction occurring when course materials are provided with CMC is said to be in line with the principles of contemporary learning strategies (Du & Wagner, 2005; Hoare & Race, 1991). As a result students are able to develop skills in critical thinking, which requires the ability to reflect on the interaction with the aid of the record storage provided by CMC technologies. While self-reflection is possible in FTF setting as well, it is usually more difficult to do in the absence of good note taking with FTF context than it is in CMC, because note taking is flawed with inaccuracies due to failed memory recall. Robson (2000) puts things into perspective when arguing that in the past, students achieved cognitive learning by linear texts and discussion. However, with technology, students are able to seek information in order to extend their knowledge, reflect on information, and then, develop their own views of the information and the foundation upon which the information is based. Du and Wagner (2005) on Weblogs refer to power of CMC to prevent free-riding that often occurs in FTF groups and other collaborative learning because assessment is graded on individual effort. Students have to take ownership of their individual contribution knowing that it will be evaluated in direct comparison with their peers.

**GIVING STUDENTS THE CONTROL**

Incorporating CMC into a course allows students to take charge of how they learn. The interaction with technology provides some level of flexibility in how students adapt technology to meet their individual needs. Olaniran (2004) argues that the ability and the freedom to ask the instructor questions directly show that CMC has a tendency to foster a greater sense of control in how they learn and in the process of learning. For instance, students do not have to wait in line to talk to the instructor or wait for appointment. The significance is that the traditional gate-keeping role of a secretary is reduced. The notion of “direct manipulation” (Fong & Kwan, 2003) speaks to the idea, where a task as simple as icon dragging or cursor position on a computer screen provide students the knowledge about the technology and also about the task. Indirectly, Rubens et al. (2005) address control when discussing the flexibility that technology design affords for knowledge building. By knowledge building, students instead of passive participation in class discussion are individually responsible for identifying ideas and to expand their knowledge base on those ideas.

Furthermore, Olaniran (2004) finds that increased control in access to instructors with CMC provides students opportunity to share more personal information with the instructor, which further aids learning as students are able to develop a rapport and comfort with an instructor, course, and the technology. On the contrary, traditional FTF interaction limits time for sharing such information. At the same time, a balance needs to be struck between offering users or students the control to accomplish a task versus control that can be counter productive. Fong and Kwan (2003) offer the following advice for effective electronic learning that users need to have technologies that provide them with knowledge of getting their work done without destroying their data. While technology is supposed to provide a means for accomplishing a task in both an effective and efficient manner, they are known to be problematic either through mistake or malfunctioning. Thus, it is advised that technology offer tool box or alert box (i.e., wizards) that can notify students of impending mistakes such as accidental deletion of document (Fong & Wang, 2003).

**COURSE MANAGEMENT**

The goal of any course regardless of the communication medium (technological or traditional) is
to help students build knowledge and demonstrate mastery of the concepts. Thus, it is important that instructors use technologies in a way that facilitates this process by selecting CMC hardware and software that create the context where students are actively engaged in the knowledge building and overall learning process. The emphasis on individualized or personalized instruction can enable students to learn, however, a mismanagement of the process can be counterproductive and inhibit learning. For instance, technology media can serve as a teaching mechanism in which case a good interface between the technology and students is required. Students must not be left to fend for themselves (e.g., learning to use the technology features). Robson (2000) hints at this when she suggests that teachers’ roles call for them to manage rather than mere pointing students to information. As instructors perform their managerial duties, attention must be given to how students learn, process information, and select technology that can help accomplish these goals. For example, in CMC using computer conferencing or threaded discussion, students need to know that they do not have to depend only on the instructor or their own ability to obtain answers to questions (Olaniran, 2004). Students can post their questions on the discussion board and ask others for help. Then instructors can monitor the discussion to mine information for areas of course or learning that instructor can offer assistance and foster deeper knowledge.

As managers, instructors can input messages to keep discussions focused on task goals and to prevent disruptive students from taking over class interaction. The CMC/FTF course design can allow instructors to identify concepts from the electronic discussion that are suitable for further FTF discussion. Similarly, tools offered within CMC media should allow for students to focus on learning rather than housekeeping tasks. Features such as auto save, auto archiving, and Wizards are tools that instructor can use to assist students to focus on learning rather than technical component that can frustrate students experience and learning. The degree which CMC technologies incorporate course management tools will influence the quality of the course and ultimately students’ learning and experience.

Instructors must decide ahead of time the degree to which a course needs to facilitate relational development among students. Therefore, it may be necessary for CMC users to establish relationships with one another and gain confidence in the system before proceeding in its use. The need for relational development in student-to-student interaction reinforces the need for supplementing CMC design courses with the FTF interaction (Olaniran, 2004; Wilson & Whitelock, 1998).

Another issue of course management in CMC courses is the degree of shift in learning control and responsibilities. CMC can shift the burden of learning from teachers to others such as students and other experts, such that teachers are able to concentrate on other aspects of learning. For instance, Olaniran (2004) finds a considerable drop in the proportion of direct teacher-student FTF interaction in comparison to the traditional FTF course. However, the reduction in time spent in FTF interactions was supplanted by increase in time needed for course management because instructors had to spend more time reading students contributions and providing individualized feedbacks. Specifically, the study reported increase in the proportion of “students contributing to in-class discussions on a consistent basis from 56 percent for FTF to 79 percent in CMC” (Olaniran, 2004, p. 152).

DESIGNING CMC INTO COURSES

A major challenge facing institutions and instructors today is not whether the CMC technology works. The challenge is making the technology work within different pedagogical goals in a way that accomplishes the idiosyncratic needs of different instructors and disciplinary demands. This
is where CMC teachers, vendors, and software designers come in and can make a difference.

One of the major benefits of any communication technology in a learning environment is to increase interactivity or active participation among students. But more importantly, CMC technologies must offer students a way to apply and extend course contents beyond the confines of their classrooms. CMC, if used appropriately, can foster critical thinking along with greater interactivity with other students (Olaniran, et al., 1996). By implication, it appears that students are expected to take initiative not only in their classrooms but also in their socio-cultural environments. Thus, educators are increasingly demanding more out of their students with the expectation that students will become active cultural change agents who are critical consumers of information (McComb, 1994). Contemporary education must find a way to engage learners’ experiences, concerns, and voices (Alavi, 1994; Du & Wagner, 2005; Freire, 1983; Leidner & Jarvenpaa, 1995). This is a paradigmatic shift from the traditional mode of education that stresses passive reaction to learning contents to the one that emphasizes active or proactive creators and participants in how they learn (McComb, 1994; Olaniran, 2004; Olaniran, et al., 1996). CMC systems designers must find a way to create suites of hardware and software that help accomplish the following:

- Allows students to connect with other learners’ experiences
- Puts teachers in a position to study and understand how best to adapt technology to students experiences in a way that leads to learning
- Allows learning to take place outside the linear mode of instructors’ transmitting or being the sole provider of knowledge
- Allows technology to extend the boundaries of learning

CMC is not a cure-all but it can offer instructors the opportunity to turn over learning responsibilities to the direct control of learners. Olaniran (2004) argues that simply adding CMC to a course will be counterproductive, however, when it is used to facilitate course goals, then it will make a significant contribution to learning processes and students, as well as instructors, would benefit from its deployment. McComb (1994) concurs when she suggests that CMC should not be used as a simple replacement for FTF communication. Other scholars have argued that using CMC in a course design does not mean that students will take advantage of it by taking the required initiatives (Olaniran, et al., 2000; Witmer, 1998). Quite the contrary, as Olaniran (2004) reported an initial indifference in communication course among graduate students who are expected to show greater latitude in taking responsibility for how they learn than the undergraduate students. Students’ resistance to technology is to be expected but this can be overcome by instructors. One way of making sure that initial resistance is overcome is by making sure that a significant portion of course evaluation is tied to performance and usage of the system (Olaniran, 1995, 2004). Furthermore, CMC can be explored or promoted by teachers from the standpoint that it offers students opportunity to allow their voices to be heard and consequently, it is useful for curricula that aim to facilitate students’ involvement and in encouraging them to develop initiatives (Du & Wagner, 2005; McComb, 1994).

FEEDBACK AND ASSIGNMENT TURNAROUND

CMC facilitates direct feedback without having to go through traditional gate-keeping roles of secretaries. However, affecting feedback is the issue of assignment turnaround. While CMC allows for assignment to be electronically submitted and graded, it is important for instructors to
understand that they would have to develop the habit of turning assignments and projects around expeditiously. Students can become impatient and frustrated when their assignments are not returned in a timely manner. The resulting frustration can also lead to failure to continue to use the technology (Olaniran, 2004). Thus, it is imperative that instructors understand that rules that work in other classes where CMC is not in use may fail miserably when applied in the CMC course design.

Another challenge with feedback in CMC lies in the fact that CMC is immune to large class size when participation is concerned. Accompanying large class sizes is also the need to grade a large number of assignments, yet the demand for fast turnaround time persists. One way to overcome this is to develop an electronic toolkit consisting of a glossary of commonly found errors, which consists of pre-written short text files about different errors (McComb, 1994). When an error is encountered, an instructor can quickly insert the accompanying text comments in student papers. First the short text files allow standardization of certain feedback on papers while allowing flexibility for personalized feedback. Second, the process allows instructors to meet the goal of fast turnaround. Third, the process allows instructor to move quickly through large volume of papers with relative ease. Fourth, the process safeguards against bad handwriting that may defeat the original purpose of feedback.

In the process of feedback, McComb (1994) considered the record-keeping feature of CMC as a useful tool for training students from comparison standpoint. Specifically, she indicates that she can refer to her notes to see if students follow her suggestions. The record also provides a good way to monitor students’ progress on activities and over a period of time. The record offers teachers the means for comparing a particular assignment with those of other groups and to present examples of poorly completed assignments against those that are exceptional (Du & Wagner, 2005; McComb, 1994; Phillips & Santoro, 1989). Furthermore, CMC allows students, when given access, to monitor their individual or collective performance on a regular basis and at their own convenience. For instance, my university provides online course management tool (i.e., Raiderlink) that instructors can use to provide students their grades and other performance evaluation measure. However, students also use it to stay on top of their grades and they use it to contact instructors about any discrepancy during the course, rather than having to wait until after the course is over or when grades have been submitted to the registrar’s office.

When designing CMC system into courses, one of the benefits is the ability to post messages and share files with participants. This becomes an invaluable tool that instructors can explore with learners especially in group interactions and other problem solving contexts. Readings, course notes, and supplemental resources can be disseminated via Internet and CMC forum. Learners can interact by sharing their comments with one another and engage in in-depth discussion with their peers and instructors. According to Bailey and Cotlar, (1994) instructors can ask different questions to stimulate interaction and thinking in CMC forum between students and multiple instructors.

**CHALLENGES AND RECOMMENDATIONS FOR CMC STRUCTURE**

A major challenge facing instructors in incorporating CMC into course design involves deciding which technology to use and when to use it. Asynchronous CMC (e.g., threaded discussion, bulletin board, e-mail) are popular because of the set up ease. The fact that asynchronous design allows participants (students and Instructors) to reflect before posting contributions is also appealing.
At the same time, the design offers the quickest solution because most existing university and IT infrastructure require little modification to add CMC to course structure. However, synchronous CMC course design (e.g., chat, video conferencing) where participants interact, at the same time, offer a greater level of interactivity and immediate feedback (Olaniran, 1994). Furthermore, synchronous interaction allows participants to interact using combination of voice, video, and audio channels, which can enhance students learning experience. With multiple channel cues’ media the barriers of time, distance, and culture can be overcome. (e.g., Solomon, 2001)

Synchronous CMC course structure could be problematic for distance learning where participants are spread across different geographic time zones. However, a recommendation for overcoming the challenge is to offer a mixed structure design, where asynchronous CMC may be the primary design. At the same time, offering synchronous interaction features such as the chat feature would significantly enhance student-to-student interactivity. Courses requiring group writing or decision making would also benefit when participants can either set up a time when all or most group members can meet to put the project together or write a paper. An alternative is for the instructor to devote class time to accommodate synchronous interactions for students, provided the infrastructure exists.

It is also recommended that software designers needs to help instructors overcome some of the challenges in CMC. One is the issue of bandwidth, which prevents against implementing more sophisticated CMC applications such as those allowing audio and video channel cues, the likes of computer videoconferencing and internet protocol television (IPTV). Creating third-party software that offers multiple cues and ease of deployment by instructors and easy access by students is called for. In the meantime, Olaniran (2004) recommends a combined method that allows for both FTF and CMC such that users do not feel like they are giving up one medium for another. Furthermore, a combined media design simulates interactions similar to those found in most organizational settings, where members are often allowed to use multiple communication media for interaction. Similarly the combined FTF and CMC design provide students time to adapt to the CMC system. Thus, while people may be using the CMC medium, they are still able to interact via FTF. The use of combined media in course design is not about CMC technology, but rather about the use of CMC to facilitate interaction and learning.

ADDRESSING STUDENTS’ NEEDS

Another challenge facing CMC and course design involves the students’ needs. Outside course content, the primary challenge facing an instructor is in accommodating different needs (e.g., learning styles and preferences) of the students. Consequently, instructors must develop a way to identify students’ needs and use the CMC system in a way that facilitates personalized learning through multiple strategies. Therefore, it is recommended that teachers break away from the linear mode of teaching that characterizes a lecture style. This can be done via prompt feedback through e-mail or posting to bulletin boards, discussion forums, along with reflective Weblogs. Another pressure placed upon instructors is that they need to be well informed by having current and adequate information to deal with individual students’ problems. Instructors can overcome this challenge by offering students “toolkits” or resource links where students can go to obtain desired information. Resources links are useful in empowering students to take control of the way they learn and recognize that learning exists beyond the classroom walls. The need for stressing individualized learning also demands some changes in procedures and course design. It is recommended that the training manual or Web
Human Computer Interaction and the Best Mix of Face-to-Face and E-Interactions in Educational Settings

Information must be easy enough for inexperienced users to understand. This simplification will help to reduce some anxieties with computer. However, the CMC training manual or Web page should not prevent sophisticated users from enjoying the system’s benefits. Olaniran (2004) advises that CMC should be used in a way that facilitates creativity. Encouraging creativity can be done when students are allowed to explore the medium for other uses that deviate from the course’s primary purpose such as socialization and recreation activities (Olaniran, 1993). A benefit from creative use of CMC is that it allows “meta-interaction” whereby students are able to explore and develop the novel use of the system as well as to seek help from one another when they encounter problems. Finally, the opportunity for students to help or coach one another in CMC forum may alleviate the need for instructors to be the only source of help.

REFERENCES


KEY TERMS

Computer-Mediated Communication (CMC): Computer-mediated communication involves communication interactions that exist over computer networks.

Critical Thinking: Involves a mental process of analyzing or evaluating information in an attempt to attain higher level of understanding. Person or learner is what has been learnt.

E-Learning: involves the process of knowledge dissemination and acquisition taken place over electronic networks.

Electronic Toolkit: represents wizard-like database that can be invoke in electronic communication by both learners and instructors to provide general but customized feedback.

Human Computer Interaction (HCI): Involves the study of interaction between people and computers where the interface of computer and people enables goal accomplishment.

Globalization: Involves economic and sociocultural ideas where organizations are able transcend national geographic and cultural boundaries through convergence of space and time in attempt to accomplish goals.

Transformational Learning: Involves knowledge acquisition that individuals and students can adapt to transform old knowledge about a given idea or topic. In other words it represents a process of getting beyond factual knowledge alone.

This work was previously published in Handbook of Research on Computer Mediated Communication, edited by S. Kelsey and K. St.Amant, pp. 49-61, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 4.2
Human Computer Interaction for Computer-Based Classroom Teaching

Wolfgang Hürst
Albert-Ludwigs-Universität Freiburg, Germany

Khaireel A. Mohamed
Albert-Ludwigs-Universität Freiburg, Germany

ABSTRACT

This chapter focuses on HCI aspects to overcome problems arising from technologies and applications that may hinder the normal teaching process in ICT-ready classrooms. It investigates different input devices on their usage and interactivity for classroom teaching and argues that pen-based computing is the mode of choice for lecturing in modern lecture halls. It also discusses the software design of the interface where digital ink, as a “first class” data type is used to communicate visual contents and interact with the ICT.

INTRODUCTION

Utilizing information and communication technology (ICT) in modern classrooms for the purpose of teaching offers several advantages. These include (worldwide) access to digital materials during lectures, recording and archiving of both the presented materials as well as the whole live event, and transmitting the lecture (live or recorded) to remote locations—a concept of tele-teaching. Consequently, conventional equipment, such as blackboards and chalk as well as overhead projectors and transparencies, have almost disappeared from the classrooms and lecture halls at many institutions and have been replaced with computers and data projectors. Integrating ICT in the classroom raises a myriad of technical questions related to the implementation of the respective devices and services. It is generally agreed, that details of the technical installation should be hidden from the presenters and teachers. Thus, in an ideal scenario, the teacher walks into the lecture room and, if anything, just has to
press one button in order to switch on the whole equipment and having all the infrastructure readily available. In this chapter, we are concentrating on the user’s perspective once all equipment is up and running and operated by the teacher during the lecture. In doing so, we focus on questions related to human computer interaction (HCI): How can people access the functionality offered by the installed ICT? How do they interact with the available infrastructure? What are the best possible approaches for the interface design in order to support the teaching process? And so on.

Generally, the way in which people interact with any kind of ICT is influenced by two aspects. The first aspect is the input devices, that is, the hardware that specifies what information is exchanged between man and machine and how this exchange takes place. For example, by moving a mouse, a user sends “events” to a computer that encodes information such as relative coordinates representing mouse movements. The second aspect is the software design of the user interface and the way in which the technical signals exchanged by the hardware are interpreted. For example, incoming signals from mouse movements (representing relative coordinates) are generally mapped to absolute pointer movements on the screen and as such visualized to the user. Unfortunately, traditional HCI concepts, such as conventional “desktop” input devices (the keyboard and mouse) and their related classical interface designs are not well suited for the tasks appearing in a classroom, forcing teachers to modify and adapt their teaching styles to the used equipment.

Following the point of view described before, this chapter is organized into two parts. First, we address input devices, that is, hardware-related issues for operating and using ICT in classrooms. We describe current and future trends in the usage of different devices using examples and case studies from our past and ongoing research projects as well as our experiences in actual teaching scenarios. We identify a subtle, yet promising trend that can be observed developing in the educational world—more and more pen-based input devices are used in classrooms and lecture halls. Examples of such devices include interactive electronic whiteboards, graphics tablets, Tablet PCs, and personal digital assistants. The second part of this chapter is devoted to the software designs and interaction paradigms for interactive classroom teaching using ICT. We describe different aspects in relation to pen-based interaction with the installed ICT. “Digital ink” is described as a new, “first-class” data type serving both, an exchange of information between teacher and students, as well as a new metaphor for interacting with the installed ICT via newly introduced gesture commands. Generally, we concentrate on instructional talks and presentations, where a presenter (i.e., the teacher) faces an audience (i.e., the students) and presents some content in a typical classroom setting. More interactive classroom settings, such as seminars or other arrangements involving group work, are discussed in a concluding future trends section at the end of this chapter.

HARDWARE-RELATED ISSUES: INPUT DEVICES

Background: Ubiquitous Computing Environments

As mentioned in the introduction, ICT usage in classrooms offers great opportunities to support and improve teaching. However, traditional input and output devices often prove to be a hindrance for the direct interaction between teacher and students. They draw too much of the lecturer’s attention to the operation of the respective hardware. Keyboard and mouse cannot be used to provide content to the students as naturally as chalk is used to write on blackboards. Monitors, even the small ones from laptop computers, can stand between the audience and the teacher building a
virtual barrier, and so on.

In a perfect scenario, all ICT equipment should be seamlessly integrated into the environment, that is, the classroom, and not interfere with the activity of teaching but support the lecturer in the best possible way. Presenters must be able to naturally access and handle the additional functionality provided by the respective devices alongside in a way that enables them to fully concentrate on their teaching and the interaction with the audience. Hence, the appropriate usage of ICT in classrooms is an ideal example for a concept of computing introduced by Marc Weiser in 1991 as ubiquitous computing (Weiser, 1991). Generally, the term ubiquitous computing refers to “an environment in which a large number of computers, ranging from hundreds of palm-sized to a few wall-sized units will be seamlessly integrated into our immediate surroundings, connected by wireless networks” (MacIntyre & Feiner, 1996, p. 251). While originally being applied mainly in the academic world, we now have the technology available to make Weiser’s vision of a ubiquitous computing environment a reality by building a classroom where all computing functionality is integrated into the environment. Many universities are already equipped with wireless networks. Large touch sensitive displays for integration into classrooms are available at prices affordable by most universities, and so forth.

However, to really create an “intelligent” classroom that does not constrain teachers but fully supports them, some open problems and challenges still remain, especially related to HCI. In the following, we give a historical overview about the usage of computers and different input devices in the classroom over the last couple of years. We discuss our experiences gained from utilizing such devices at our university and from observing other project- and user-groups. The aims of this overview are to highlight the advantages and disadvantages of the respective devices, identify interaction and usage styles, provide advice for the appropriate usage of the technology, and pinpoint open problems for further HCI-related research. Pen-based interfaces are identified as a trend-setting and essential aspect for ICT-based classroom teaching. Afterward, two case studies of innovative usage of pen-based input-devices are presented, which give direction to the future usage of pen-based interaction in an interactive classroom.

**Input Devices for Computer Usage in Classrooms: An Historical Overview**

*Traditional teaching: blackboards, overhead projectors, and freehand writings.* Teaching a class involves a lot of interaction between teacher and students, through the freehand writings on blackboards or chalkboards. Blackboards have traditionally been by far the most widely used medium for lecture presentations in institutes of higher learning, and they are very effective. Not only do they allow instructors to communicate all or part of the course materials in natural handwriting, but they also take away the attention to details for their use. Freehand writings and annotations take an important role in classroom teaching, even if a lesson is delivered via previously prepared slides or transparencies that are presented to the students using an overhead projector. Essentially, as classes progress, teachers will write, highlight, underline, draw, and formulate on those slides to put forth their points to the students. By using “empty” slides, they can spontaneously write on them, what actually facilitates discussions amongst students, helps to illustrate complex mathematical proofs, and demonstrates diagrammatic facts more clearly. Studies show that the combination of handwritten annotations, the development of thoughts through freehand writings, and the verbal explanations of the lecturers, are indeed crucial in assisting students’ understanding of the subjects (Mukhopadhyay & Smith, 1999; Zupancic & Horz, 2002).

*Early computer usage: data projectors and presentation software.* With the rising popularity
of presentation software tools, such as Microsoft’s PowerPoint, in the mid and late 1990s, a sizable number of teachers moved from using chalk and blackboards or overhead transparencies to using electronic slides and computers for classroom teaching. The increase of processing power and the decrease in hardware pricing led to the augmentation of lecture halls and classrooms with data projectors connected to background computer systems. By observing how teachers took to using these new tools, three different trends can be discovered:

1. Due to the lack of a flexible and effective input device such as the pen and partly due to a less than perfect interface design, some lecturers adapted from their very lively teaching style, to a scenario where the materials presented to the students were prepared electronically in advance and shown in a pure “slide show” style presentation. Hence our observation that whenever computers are exploited in classrooms, the teaching styles of the instructors are often made to adapt to the currently available interfaces and not vice versa.

2. Some teachers have tried to revert back to their old, traditional approach of slowly developing a topic by continuously adding and annotating their contents. This often results in their sitting down behind a computer screen and using traditional input devices, such as the keyboard and mouse, to add and modify the presented materials during the lecture. Where instead of writing in freehand and annotating on the slides, they type the text with the keyboard and make annotations by drawing straight lines and boxes with the mouse. So, again, the teachers go off their traditional way of teaching and adapting to a new and less appropriate style of interaction. In addition, sitting down and “hiding” behind a big computer monitor led to a loss of contact with the students.

3. Chong and Sakauchi (2000) observed that when presentation slides are available (either as overhead projected slides or PowerPoint slides from a data projector), the use of the chalkboard is significantly reduced during the course of the lecture. However, when it comes to facilitating questions and answers, the chalkboard is conveniently used again to demonstrate facts, confirm equations, and diagrammatically illustrate flowcharts. But making a direct reference from the chalkboards to any of the slides can be a hassle, and sometimes the instructors resort to annotating on those slides in whatever little space is available, and may even compromise on legibility.

In all three scenarios, we can observe that forced by inappropriate user interfaces, teachers generally adapt their teaching style in a less preferable way to the available technology. As for the students, they normally appreciate the fact of having an electronic version of the presented material after the lecture. It enables them to pay more attention to the class, and it gives them the time to make more appropriate and “intelligent” annotations during the live event because they do not have to copy everything the presenter shows them. However, pure slide shows are generally considered too exhaustive. Students often complain that such presentations are too fast for their liking, while the required speed for handwriting naturally slows down the pace of the lecture to a speed that is generally considered more appropriate and pleasant.

Initial usage of pen-based devices: Graphics tablets and touch screens. Considering the negative observations described previously, it is no surprise that with the advent of devices for pen-based computing, presenters quickly started to experiment with them in their classrooms and lecture halls. People first started using touch screens (where users communicate with the computer by direct touching of the screen) and
digitizing or graphics tablets (where a digital pen is moved over a separate surface in order to remotely control and operate the computer, similar to moving a mouse in order to operate the mouse pointer on the screen, cf. Figure 1). These input modalities improved the situation in so far as teachers are at least able to write natural freehand annotations. The devices, however, still lacked the usability and simplicity of markers or chalk. For example, in the case of graphics tablets, not every teacher feels comfortable with the remote control of a digital pen, but prefers to write directly onto the visualized electronic slide. Touch screens are hard to write on when the monitor is standing upright. Ideally, the screen should be mounted horizontally. However, touch screens do not only react to pen input but to any kind of physical contacts, thus making it impossible for teachers to rest their hand on the surface while writing. A detailed discussion about the initial usage of pen-based input devices in the classroom can be found in [*].

Advanced pen-based input devices: Interactive LCD panels and electronic whiteboards. The previously described input devices represented an improvement over traditional ways for human computer interaction such as keyboard and mouse. However, it was only after large, blackboard-like electronic whiteboards and LCD panels with integrated tablet became available at a reasonable price when pen-based computing became a real alternative to chalk and blackboard usage.

Originally targeted toward special interest groups and professions (such as graphic designers), the first LCD displays with integrated tablet, in the following called interactive LCD panels (cf. Figure 2), became available for the average consumer by the end of the 1990s. These devices operate like regular graphics tablets described before with the main difference being that the tablet is transparent and mounted on top of an LCD panel. Hence, compared to common touch screens, the surface does not react to any kind of contact, but only to the input of a special pen, thus enabling presenters to naturally write on a horizontally mounted screen. Using the pen, people can directly interact with the applications (e.g., by making annotations on the presented slides, navigating through subsequent slides, etc.). Teachers generally see this as a big improvement compared to normal touch screens or graphics tablets. However, early versions often proved to be too small, had a resolution that was too low, and a limited processing power that resulted in a noticeable time delay during freehand writing. Nowadays, we observe significant advancements.
in display size, screen resolution, as well as processing power, paired with lower prices. Hence, when mounted horizontally such devices provide the same ease of use as traditional overhead projectors while at the same time enabling users to access and use the full functionality offered by the underlying ICT installation.

An electronic whiteboard (cf. Figure 3) is the equivalent of a chalkboard, but on a large, wall-mounted, touch-sensitive digital screen that is connected to a data projector and a computer (see, for example, Smart Technologies Inc. [Online]). This computer can be controlled directly from the digital screen with a digital pen or by simply using a finger to touch it. Other known terminologies of the electronic whiteboards include the “eBoard,” “digital whiteboard,” “smart board,” or “interactive whiteboard.” They carry slightly different meanings to different people depending on their environment of application. However, these terms all describe in general, the group of technologies that are brought together to support classroom activities. Different versions exist that rely on front or rear projection. In the case of front projection, there are also whiteboard solutions that do not react to any kind of contact, but (similar to interactive LCD panels) use a special pen for interaction. While such an approach is preferable for LCD panels or any kind of horizontally mounted input surface, it is less important for vertically wall-mounted boards since users generally do not rest their hand on the board during writing.

Similar to early interactive LCD panels, first versions of electronic whiteboards had limitations regarding size, resolution, and processing power and were often too expensive for large-scale usage. Again, the situation is improved significantly today, and we can observe how more and more classrooms are becoming equipped with such devices. In fact, there is a growing trend amongst learning institutions approving the integration of electronic whiteboards into their classrooms. There are also many reports in the emerging body of literature suggesting that vendors of education and teaching instructors alike find the electronic whiteboards relatively easy and compelling to use (BECTA, 2003; Glover & Miller, 2001). However,
The authors also caution that becoming confident in their use often requires effort in commitment in terms of both training and independent exploration.

The situation today. As described previously, we saw that pen-based input devices offer promising perspectives for usage in classroom teaching, but initial versions had limitations and disadvantages. Due to advancements in hardware (such as size, resolution, and processing power) and dropping prices we can now say that electronic whiteboards and interactive LCD panels have become a real alternative to traditional blackboards and overhead projectors, respectively. Both devices offer a similar ease of use as their traditional counterparts while at the same time opening various new opportunities by providing full access to the underlying computer systems and infrastructure. In addition, new developments such as Tablet PCs (cf. Figure 4) incorporate the pen-based input technology and are more and more used for giving presentations and classroom lectures as well. Tablet PCs are a special kind of computer in the size of a laptop with a pen-operated graphics tablet integrated into the screen, similar to the interactive LCD panels described previously. Some Tablet PCs purely rely on pen-based user interaction, while others, called **convertibles** or **hybrids**, offer pen-based operation in addition to a regular keyboard and mouse or touch pad input.

Based on the feedback we received in various situations at our university, we argue that the equipment available today can finally be used to replace blackboards and overhead projectors through their electronic counterparts in a reasonable way. We observed that even people who are rather skeptical about using computers in classrooms and still stood with the old, traditional form of classroom teaching are more and more switching to computer supported presentations due to these new input devices. Overall, we can conclude...
that pen-based computing will be an important if not essential part of computer usage in future classrooms. It seems likely that in the near future, more and more classrooms will be equipped with pen-based input devices, in a similar way as not so long ago, blackboards and overhead projectors were complemented and slowly replaced with electronic data projectors.

In the following section, we describe two case studies of actual equipment used at our university for teaching. The purpose of this description is to describe the respective technology and how it is best used, to illustrate how people take advantage of it, and to identify open problems and future trends. We will see that with this technology we are able to build real ubiquitous computing environments as described before as the ultimate goal in the “Background” section.

**Case Studies of Actual Equipment Used in a Ubiquitous Computing Environment**

*Electronic whiteboards as blackboard replacement.* As mentioned before, today’s electronic whiteboards are larger in size, offer a higher resolution, are technically more reliable, and are easier to handle. However, their absolute size still remains an issue when compared to traditional blackboards. As a consequence, some people started placing two boards next to each other in order to increase their writing space by forming a large electronic whiteboard (Prante, Streitz, & Tandler, 2004). Examples of such “twofold” whiteboard installations at our university can be found in Figure 5. The boards shown at the top and bottom left of Figure 5 are integrated into the front wall of the lecture room thus hiding the required technology from the presenter and providing an environment, where the computing functionality can be directly accessed by the teacher through interacting with the electronic whiteboards, but the...
whole technology does not interfere with regular teaching. All technologies are hidden as much as possible in order to let the presenter concentrate on the talk without distraction from any of the hardware. In our installation, both boards can be connected to one computer. People are able to place one application window (e.g. from the used presentation software) over two boards thus creating one large “virtual” electronic whiteboard. Experience with users has shown that the border in the middle of these two boards is generally not considered critical since most presenters prefer to write on the board in columns anyway. While connecting several input and output devices to one computer was still considered critical a few years ago (MacIntyre & Feiner, 1996), today’s windowing systems generally support this kind of functionality without any problems.

When observing how people are using such an installation, we realized different approaches. Some presenters like to have just one application window that extends over both whiteboards, either using it to present two slides next to each other or to start with two empty slides and an increased writing surface. Others prefer to have two separate applications running, for example, in order to show an overview slide of the lecture on the left board while presenting the current slide on the right side. Some teachers use one board for slides and the other one to show a different application, such as a player for video clips or animations. People can also connect their own laptop to one of the boards, which is used, for example, by some presenters to show their slides from the laptop, while referring to the other board (and the connected computer in the classroom) when accessing and presenting information from the Internet. Using the two output devices for different applications is a common practice often chosen when using multiple monitors for desktop computing as, for example, Grudin’s studies pointed out; there, participants used multiple monitors rather for partitioning than to increase space (Grudin, 2001).

Overall, we can observe that people take great advantage of the additional working space. No single trend in using this installation can be ob-

Figure 5. Putting two electronic whiteboards next to each other, in order to form a larger input and output area, resemble traditional blackboards.
served, but a variety of different usage scenarios exists. Hence, instead of having the teachers adapt to a certain interface or style of interaction, this setup enables them to choose among different approaches and to use the respective tools in the way they personally prefer.

*Integrated interactive LCD panel as overhead projector replacement.* In addition to the boards, we built a high desk or lectern with integrated computing facilities that can serve as a true replacement of the traditional overhead projector (cf. Figure 6). As input device, a horizontally mounted LCD display with pen-based interaction functionality is integrated into the surface of the lectern. All other equipment needed for a talk is embedded into the lectern, but again hidden to the user as much as possible. This includes even audio recording facilities if the presenter wants to record the lecture with some lecture recording software (Müller & Ottmann, 2000) and preserve it for further, off-line usage. Hence, the normal user just has to connect the lectern to the power source and data projector and press a button to turn it on. Including all required equipment into the lectern provides several advantages, because it enables us to make the whole device mobile by putting it on wheels. Thus, it can be easily maintained and replaced by technical staff if necessary. In addition, it can be transferred to different rooms, which is important in case not all rooms can be provided with the full range of technical equipment. It also allows one to easily put it away and lock it, in case it is not possible to always lock the room.

Considering the usage by the presenters, most of them reported that handwriting and making annotations is much easier, since they can do this on a horizontally mounted surface placed in front of them. Some also expressed their appreciation that they can face their material and the audience at the same time and do not have to continuously turn their back to the students when presenting their material. Meanwhile, several of these lecterns were built and used in various lecture halls at our as well as other institutions.

Both the lectern as well as the large interactive whiteboard described before turned out to be reasonable replacements for traditional devices used

*Figure 6. Lectern with integrated computing facilities and pen-based input device. The interactive LCD panel is integrated into the surface in order to support freehand writing on the presented slides and pen-based operation of the used applications.*
in classrooms, that is, blackboards and overhead projectors. If integrated into the environment as shown, for example, in Figure 5 and 6, they do not interfere with the normal teaching process of the presenter while at the same time offering full access to the computing functionality offered by the underlying computing system and network infrastructure.

SOFTWARE-RELATED ISSUES: INTERACTION PARADIGM

Background: Digital Ink

In the previous section we argued that the pen is the natural choice for interaction with wall-mounted boards and interactive tablets. It enables users to interact directly with the respective tools instead of remotely as with a mouse pointer. Freestanding writing proves to be more intuitive and flexible to enter content during a lecture than a keyboard. Hence, handwriting becomes a “first-class” data type, on an equal footing with text entered via keyboards. However, common interface designs and graphical user interfaces (GUIs) usually follow the desktop metaphor and are therefore optimized for input devices such as keyboard and mouse. For example, navigation through nested menus can easily be done with a mouse, but is hard to do with a pen. For pen-based interaction, earlier research on pie menus proved an advantage for users of digital screens, as people tend to better remember the cyclic position, and as such expedites the selection process (Callahan, Hopkins, Weiser, & Shneiderman, 1988). Hence, we see ourselves confronted with two different demands: on the one hand, the aim of using pen-based input devices in order to present information to the students, while on the other hand, the need to rely on traditional, mouse-based input in order to interact with the ICT and to operate the respective programs. To avoid the unfavorable switching between different devices, we propose a new interaction mode purely based on pen input and the notion of digital ink described in the following sections.

Digital Ink. Motivated by the treatment of handwriting as a “first-class” data type, we introduce the notion of digital ink, which is used not only to enter and represent contents, but also to operate the ICT via gesture commands. A gesture command is the resultant process of invoking gesture-like movements with the pen, that resemble special shapes drawn in a certain distinctive way, recognized by the program as certain predefined instructions for it to carry out. The digital ink derives its name from the data class that represents any kind of information created when using a digital pen to draw strokes on a screen of a pen-based input device. While the strokes received on the pen-computer’s surface get recorded as digital ink, its consequential markings are reflected noticeably onto the screen simultaneously. Any digital ink that was written is either kept in its raw form or goes through a process that translates it into recognizable texts or graphic objects. Often, in many pen-based applications, the ink goes further to be tried as a possible pen-gesture. In this case, when the ink is confirmed to be a gesture command, its visual trace on the screen is removed upon the execution of the respective gesture command.

Digital ink in its raw representation as processed by the respective hardware and software holds far greater amounts of information than are immediately apparent. These include, among others, the type of pen tip, the amount of pressure used to create a stroke, the height of the pen above the screen surface, opacity value, color, and timing data. As a result, the digital ink is acknowledged in serious enterprises such as one for secure handwritten signatures (Cyber SIGN, Inc., 2005; Gries, 2000; Hangai, Yamanaka, & Hamamoto, 2000), as well as in everyday norms of handwriting e-mail messages (Pen&Internet® LLC, 2005), sketching design posters (Corel™, 2005), and annotating digital documents (GRAHL...
As more electronic devices with pen interfaces continue to become available for entering and manipulating information, efforts have been made to ensure that supporting pen-applications are effective at leveraging this method of input. Handwriting is an input modality that is very familiar for most people since everyone learns to write in school. So, there is a high tendency for people to use this mode of input and control not only in the classroom scenario described previously, but for a great variety of different applications as well.

**Ink traces and symbolic representation.** The domain of the digital ink is unlike any others, such as text and graphics, that can have all its related information easily organized and cross-referenced, and be presented in front of users to allow them direct manipulation of the data. A trace refers to a trail of digital ink data made between a successive pair of pen-down and pen-up events representing a sequence of contiguous ink points—the $X$ and $Y$ coordinates of the pen’s position. If the sampling property of the input device is not constant, it becomes advantageous to include timestamps for each pair of the sampled coordinates. A sequence of traces symbolically accumulates to meaningful graphics, forming what we humans perceive as characters, words, drawings, or gestures representing commands. Each trace can be categorized in terms of the timings noted for its duration, lead, and lag times.

The importance of digital ink has also been recognized by the Internet community, leading to the specification of the World Wide Web Consortium’s Ink-Markup-Language (InkML) standard (Russel, Chee, Seni, Yaeger, Tremblay, Franke, Madhvanath, & Froumentin, 2004). InkML allows the description of precise digital ink information that is not only interchangeable between systems, but also allows the convenient “archival” and “streaming” of ink data between applications. Among many of its properties, it is important to note that:

- InkML provides a common format for the exchange of ink data between components such as handwriting and gesture recognizers, signature verifiers, and ink-aware modules;
- InkML supports the representation of hand-drawn ink with full attributed information, on top of capturing the pen positions over time;
- InkML allows the specification of extra information for additional accuracy such as pen tilt and pen tip force (pressure) for the enhanced support of applications such as character and handwriting-style recognition, and authentication;
- InkML provides the means for extension—by allowing application-specific information to be added to the ink files to support issues of compatibility.

**Active digital ink in recordings and replay.** One of the advantages of the usage of ICT in classrooms is the possibility to record and archive the live event for further usage and studying. Today, numerous systems exist for automatic lecture recording that automatically turn the live event into a multimedia document. These files are used, for example, by students to reflect a lecture. Others use them to deepen their understanding of particular topics. They can even be used as the core for building Web-based online courses.

Independent of the respective application scenario, the studies mentioned previously (Mukhopadhyay & Smith, 1999; Zupancic & Horz, 2002) highlight the importance of the development of thoughts through freehand writings. They also confirm the evidence that simply capturing the static end-result of the lecturer’s annotations for recalling the subject at a later time is inferior to unambiguously maintaining the whole sequence and dynamics of those freehand annotations when recorded for interactive replay.

As a consequence, user interfaces that allow replaying of real-time multimedia documents...
usually include command buttons to quickly “fast-forward” and “rewind” their data sets for systematic browsing. We see this very often in many learning applications that try to cater to both students and teachers in assisting them to effectively retrieve desired segments of those documents. Random access navigation, that is, the ability to access any position along the time-line by using, for example, a slider interface, provides an additional advantage on top of this simple fast-forward and rewinding of real-time information. It gives users the ability to straight away focus on the intended portions of their documents without having to wait a second longer for these processes to arrive at the stipulated point of time, albeit the processes being a little faster than the normal replay mechanism. Hence, users should be able to navigate through digital ink recordings along the timeline. However, realizing such a random visible scrolling is a challenging task.

Summary. In this section, we introduced digital ink as a notion to describe pen-based input, which is used not only to model chalk or markers in the digital world, but also to interpret particular ink traces, which are the building blocks of digital ink, as gesture commands. In such a setup, we need sophisticated techniques, not only to represent, process, and manipulate traces of ink, but also to store, access, and replay them. The ability for random access navigation poses a particular challenge in this context.

In the following section, we first describe aspects about the personalization, archival, and streaming of digital ink. Subsequently, we introduce a new paradigm for interaction with electronic whiteboards via ink traces that are interpreted as gesture commands.

Personalizing Digital Ink for Archival and Streaming

In addressing the proper support interfaces for digital ink mediums that are to be stored and indexed in accessible archives or streamed in broadcast networks, appropriate tools are required for practical personal manipulation of the digital ink within multimedia and multivalent materials (Phelps & Wilensky, 1997). These tools are for the producing, capturing, and replaying of freehand writings, and should therefore contain features deemed to let users feel that their digital handwritings are acceptable substitutes for their normal ones when applied on digital documents. We have identified six main feature-themes that relate to what users perceive in the contents of the formulated digital strokes and what they would expect when interfacing with them. These feature-themes refer to the personal flavor, manipulation in pixels, manipulation by objects, compactness of representation and scalability, replay and synchronization, and facilities for random-access.

1. Personal flavor: Every individual person has a unique and particular writing style. This is portrayed in the calligraphic output characteristics from their written works. A handwriting tool needs to allow the expression of this style, as accurately as possible. The recording samples and the replay stream of this tool should maintain all these personal calligraphic attributes. Various pen tip styles, such as coarse whiteboard marker-pens or fine pencil leads, and pressure-sensitive information should be supported during both recording (creation) and replay. Note that most currently available digital ink software offered in presentation tools is either nothing more than fixed-size pixel-traces on screen, or does not support the capture and replay of traces in a manner that maintains the dynamics of the generation process. That being mentioned, users should also expect simple anti-aliasing techniques on the interface tool to improve the stroke appearance of the traces rendered on screen (Klassen, 1991).

2. Manipulation in pixels: Removing traces, or parts of it, and inserting empty spaces
on the grid-like digital screens are typical examples of pixel-oriented operations.

3. **Manipulation in objects:** Generating or selecting whole objects, stretching, contracting, reflecting, pivoting, copying, and pasting are typical examples of object-oriented operations applied to trace-objects. Arguably, the typical editing commands in modern graphics tools (insert, copy, paste, and the like), however, may also be considered as pixel-based operations.

4. **Compactness of representation and scalability:** Extensive writing ultimately generates large volumes of data if the complete transcript of the raw digital ink is stored in an uncompressed form. Inevitably, eliminating any superfluous information will be a reasonable course of action. One way to ensure that the integrity of the data remains intact would be to try and symbolically represent the handwritten letters/alphabets as spline-curves that are independent of the screen resolution. That is, scaling such spline-curve-writings up or down would preserve and not deteriorate the rendering quality of the letters/alphabets during replay of the recorded traces.

5. **Replay and synchronization:** At this point, it should now be possible to replay all recorded handwritings in their best quality, the same way as they were first recorded/presented. Here, both the inter- and the intra-stream time constraints should be observed, such that the full dynamics of the generation process of handwritings are maintained in exactly the same way as they appeared at generation time. Furthermore, the interface must allow all (time-stamped) data streams to be synchronized with other recorded data streams, particularly with the audio stream of the lecturer’s voice. The latter is usually considered to be the *master* stream in a master-slave synchronization model for replay (Müller & Ottmann, 2000), whereas the recorded ink traces are considered as the *slave* stream that is synchronously replayed with the master stream.

6. **Facilities for random-access:** Easy navigation, particularly “random visible scrolling” (cf. previous) through a recorded document, requires random real-time access to the entire data collection. Metaphorically, this ideology resembles the thumb-flipping of a book. Given an instance of time, the complete list of objects visible at this time instance must be made available for display almost “immediately.” That is, it should be possible to compute and display all of these objects fast enough, so that the user gets the impression of immediacy, as opposed to making the user wait due to latency issues.

Table 1 summarizes the following discussion about how current enterprise tools fulfill these six requirements.

The Web tool **riteMail** (Pen&Internet® LLC, 2005) allows for handwritten messages in an interactive, electronic ink e-mail application. Users can create high resolution, fully interactive handwritten notes and drawings on hand-held devices, Tablet PCs, as well as on every Java-enabled desktop or laptop computer. The notes can then be sent to any e-mail address. The riteMail software employs unified Extensible Markup Language (XML)-based protocols for digital ink transmission between all components of the service. However, a replay facility is not included within (and, in order to be fair, not essential for its purpose).

**Corel Grafigo™** (Corel™, 2005) targets pen-computers and is designed specially for the Windows XP Tablet PC Edition. It supports pressure-sensitive pens, allows the marking up of documents and images individually, and provides the features of shape recognition, and gesture-based commands. Data are saved in SVG format.
Painter™ Classic, marketed by Corel (Wacom, 2005), is an excellent paint program for graphic designers that simulates the various graphic markers and coloured pencils and pens, as well as water color and oil paints. As an image-editing program, it offers numerous high performance functions (edged and smooth drawing, distorting, cloning, etc.).

The AOF-tools and their commercial spin-off, Lecturnity (imc AG, 2003), fulfill many of the interface requirements discussed in the previous section, except that they do not fully preserve and regurgitate their user’s personal style of writing. Moreover, the data volume on recorded documents with extensive handwritings is intolerable. The reason is its highly redundant storage method; for while Lecturnity offers a real-time random access facility during replay, it also computes the entire list of visible objects on the screen for every single time instance.

Using Digital Ink on Electronic Whiteboards

In the preceding discussion about input devices we already described different usage scenarios of electronic whiteboards in a ubiquitous computing classroom. Now we take a closer look at the actual applications running on those devices, how people are operating them, and what kind of interaction takes place.

Instructors and students in class do not only see the prepared course materials projected onto the electronic whiteboards, but also the “active” annotations made by the instructors using the digital pens as the lessons progress. Combined with dedicated software applications, the supporting whiteboard technologies can be made to function more effectively with better pen control, and appropriately displaying both static and moving images at suitable resolutions. Among other things, these applications generally offer:

- “Flipchart” capabilities to save and print images as well as annotated ink-trace markings;
- A “blackboard” to manipulate primitive handwritten text and drawings;
- Integrated slides and blackboard “environment”;  
- “Movie screens” to showcase video and software demonstrations;
- “Tools” to facilitate discussions and group work; and
- Access to the network, where shared teaching resources may be available.

### Table 1. Summarized assessment of various tools on freehand writings

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Software</th>
<th>Writing style</th>
<th>Support of pixel-based operations</th>
<th>Object-oriented manipulation</th>
<th>Symbolic representation</th>
<th>Recording</th>
<th>Replay</th>
<th>Random access</th>
</tr>
</thead>
<tbody>
<tr>
<td>RiteMail</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>static</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Corel Grafigo</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>static</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Painter Classic</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>static</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Tablet PC Software Development Kit</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>static</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Lecturnity</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>dynamic</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>
More importantly, the way the electronic whiteboards are utilized will shape the kind of impact they will have on the overall learning environment. The basic guidelines highlighted by Glover and Miller (2001) for good practices and outcomes of engaging classroom interaction and participation with the electronic whiteboards across the curriculum are:

- To increase efficiency by enabling teachers to draw upon a variety of ICT-based resources without disruption or loss of space;
- To extend learning by using more engaging materials to explain concepts; and
- To transform learning by creating new learning styles stimulated by interaction with the electronic whiteboards.

All of the previously mentioned technological and social benefits have positive indications from the wider observational point of view. Nevertheless, there are also subtle cases that can make instructors hesitant to change their pedagogy to incorporate the electronic whiteboards into their lesson plans, if practical considerations hinder their perspectives. We already pointed out that it requires effort in commitment to develop media-enhanced teaching contents for the electronic whiteboards, which could be considerably taxing on the teachers’ initial workload. Also, according to the studies compiled by BECTA (2003), while the act of teaching with electronic whiteboards instills anxiety in the students to learn, it unduly places the pressure onto the teachers when it comes to delivering the contents. This is especially true if the writing environment of the electronic whiteboards is not sufficiently supported with the necessary interfacing tools. We have thoroughly discussed this in the preceding sections.

Good electronic whiteboard contents are generally those that can display themselves immediately in highly visual terms, which can be most effectively understood by a group of individuals (BECTA, 2003; Clyde, 2004). The ability to easily modify and change these content materials is all about the personalization of the electronic whiteboard environment. In relation to this, and in order to achieve good content presentation, there needs to be good software applications to manage and control the content data with enhanced features and functions. Currently, makers of the electronic whiteboards ship out accompanying software that is embedded into their systems for use in classroom environments. Examples include those that are listed on Smart Technologies Inc.’s Web site such as notebook whiteboarding software and SMART Board tools. These intuitive programs support numerous operations that are categorized into many menu items under several options that teachers can use while delivering their lessons. So much so, that they tend to become more of a distraction to the teachers rather than an assistance to them in front of a live audience.

Writing and “feeling in control” of the board environment. Now, we are particularly interested in the part where the previous support interfacing software allows teachers to write on the electronic whiteboard freely as if it were a normal blackboard. The situation described previously may not necessarily pose as a problem if the screen-size of the electronic whiteboard is small. However, when these programs are maximized to full screen, then teachers will bump into the problem of “hand-and-eye coordination.” This is a cross between uncomfortable peripheral views for the eyes, and awkward bodily positions to access the menus without disturbing the flow of the lesson, or blocking the audience. This problem for a single large screen worsens when we cascade the electronic boards in the learning environment in series next to each other as described in one of the previous case studies (cf. Figure 5).

Without investing in additional tracking devices such as motion-detectors or cameras, the large interconnected, high-resolution, wall-mounted, multiple displays are by themselves receptively sensitive to the teachers’ touch. Properly configured, this electronic whiteboard
environment offers an extended desktop metaphor quite unlike the ones that we presently know and have grown used to. Further incorporating the pen-input technology into this setup accords teachers a higher level of personal presentiment than a conventional monitor, in a way that it allows them to directly touch the active-screens and see immediate reactions through the digital pens. The electronic whiteboards are meant to be an active presentation medium for a congregation of audience in a large room. While there is always a desktop console that controls the electronic whiteboards, through the mouse and keyboard, freeing up this mode of interfacing lets teachers communicate directly with the screen using just a digital pen.

Judging by the dominant metaphor that what we have today is a mismatch for the computer environment we are dealing with tomorrow, it is difficult to place this huge screen scenario (alone) into any pronounced categories of user interface debates on concepts of adaptivity and adaptability (Shneiderman & Maes, 1997). For instance, the sphere of influence of the hand-and-eye coordination needs to be enlarged, and perhaps include the body, to follow the claims of direct manipulation of adaptable interfaces. Furthermore, while the audience have it easy watching from a distance, the teacher does not: standing so close to one part of the multiple huge screens often leads to the interface widgets he or she may require to be out of reach, or worse still, because the teacher cannot see those widgets, he or she may assume that such actions or commands represented by those widgets do not exist within the board application. We point out here that this may affect the flow of the delivery of the lesson contents. In this case, we may be left to rely on the adaptivity of the interface for the electronic whiteboards to proactively assist the instructor while at the same time ensuring that they still feel in control of the whole process.

**Interacting with Electronic Whiteboards Through Gestures**

In order to solve the problems previously described, we now describe an interaction concept that realizes a modeless interactivity by deriving gestures from digital ink traces. We define *modeless interactivity* for the digital ink environment on the electronic whiteboards as the seamless integration of both the “writing” and “gesture-command” modes on a common, unified platform. On top of this, they further encompass the conventional on-screen widgets such as the fixed pull-down and pop-up menus as “interaction modes” that can be minimized on their usage or done away with completely when not needed in the most cumbersome of circumstances. They go by the notion that instructors should not always need to go between different sections of the boards in order to effect a menu action.

*Deriving gestures from digital ink traces.* It is possible to convincingly anticipate, based on the lead and lag times obtained, that the latest ink component is more of an instance of a trace, rather than a gesture. Advanced recognition techniques have been introduced to the digital ink domain, and while a number of works have gone into handwriting and sketch recognition (Rubin, 1991), there are among these works that use the same techniques to attempt to recognize hand-drawn pen gestures.

Hence, the “writing” mode as well as the “gesture” mode can help out with the current interface issues. “Gesturing to the system” refers to the actual gesticulated actions resulting from the pen movements that represent a command, and not merely the “tap-and-drag-the-icon” movements. The prospects of gesturing commands in ink-domain applications are encouraging. So much so that many authors think that this is the epitome of the final interface for pen-based applications. Although it is agreeable to a certain extent, that this mode of ubiquitous interfacing may form the fundamentals of future (more enhanced) hardware
devices, we should not anticipate this to be the trend in this current context. The perceptual user interface (PUI) guidelines should be adopted and used to create the first steps of the bridging between fully packed screens of user interface widgets to that of a ubiquitous one (Shneiderman & Maes, 1997).

**Deriving interfaces from gestures.** For the same technique that a right-mouse-click can bring up a popup menu on documents that contain menu-items “generally” related to the application it serves, pen gestures have the relative capability to have this metaphor extended. They make it possible to provide more “specific” menu-commands that are currently in context. By this, we mean the placing of directed (or pointedly explicit) “context-specific” menu-items that should obviously take precedence on the front-most tier of the multi-cascading popup interface by virtue of what is currently being displayed on screen. The rest of the more “general” menu-items, however, are neatly tucked away in the lower-priority cascades to ensure that they are still accessible, as is required by the specifications of perceptual user interfaces. For example, right-clicking anywhere on a Word document (or most Windows applications for that matter) brings up menu-items generally affiliated to the Edit-menu on the application’s main menu-bar. The advantage here is that users do not have to skit their mouse over a distance to access the main menus. This same advantage is magnified many times over when working in front of a series of large wall-mounted, hi-res digital screens where the pen is the primary input modal.

The extension we mentioned previously for this technique will involve additional information we can gather from the background environment, and is designed mainly for the convenience of the wall-mounted SMART electronic board users. We discussed this thoroughly in our previous work that deals with digital whiteboard-applications for classroom scenarios, where the sole objective is to assist lecturers in delivering their lessons comfortably and not worry about the support technology.

We incorporated an active program module called the Context-Assembler, running in the background of the main program application, that keeps track of all “items” that are visible on the digital board’s writing area. These include all digital ink-entities that constitute drawings, writings, illustrations, and so forth, and other screen-related widgets like images, fonts, and the writing canvas itself. These “items” are ordered and cross-referenced with each other in an efficient way in the Context-Assembler to give optimal background support information known as “context-information.” That is, if we ask the Context-Assembler about one particular trace that we wrote earlier on the writing canvas, it will tell us when and where the trace was created and located, which other “items” are its immediate neighbors, which are the other items’ it is overlapping with, which group(s) of “paragraph” it is connected to, the probability of the trace being a gesture, and all other “items” that are similar to it in terms of appearance, size, color, and style. Based on this context-information that can be retrieved off any “items” in the Context-Assembler, we designed a “context-oriented” popup menu interface.

The mechanism that determines what menu-items receive higher priorities over others, in order to appear on the front-most tier of the cascading popup menu interface, is an application-defined mapping function that uses the context-information as its input parameters. For example, in Figure 7, the latest ink-element is queried for its context-information. Among other things, the Context-Assembler replies by telling that the ink-element in question is on top of another ink-element created at time \( (t-m) \), is on top of an image known as “slide p28,” is on top of the writing area known as “p9,” stretches horizontally across these three previous “items”—from left to right—and has a very high probability of being a “right-arrow” gesture. Putting these into the mapping function returns a list of thematically categorized menu-items that interprets the last ink-element as a command gesture. The top few
Figure 7. The latest ink-element in (a) is interpreted as a gesture that can have several interpretations. The current context of the board (writing area p9, slide p28, and ink@time(t-m)) is determined and assembled appropriately in (b). The output of the mapping function is then rendered onto the screen as support interface on-demand (c).

high-priority menu-items in the list will suggest that the command gesture advance to the next “slide,” or to the next “page,” or to next “slide” and “page,” or to move the overlapped ink-elements identified in the context-information to a new location.

To complete the visualization of the on-demand assistance, we bank-in on the last known coordinate position of that ink-element in-context. Popping up dynamic-content pie menus and/or other static command buttons, with respect to the list of menu-items returned by the mapping function, in that vicinity seems most appropriate and serves as a convenience to the user. As said before, pie menus proved an advantage for users of digital screens as earlier research pointed out (Callahan et al., 1988). Our command button (toolbar) interface is logged to the bottom of the screens, appearing on-demand with only an array of anticipated widgets. The combined lengths of the number of widgets will not exceed the arm length of the user’s to ensure that all widgets are within reach and within the user’s line of sight.

CONCLUSION AND FUTURE TRENDS

Freehand writing is a natural way of exchanging viewpoints and communicating visual ideas between teachers and students in a classroom. Motivated by this, we identified the value of pen-based computing for the usage of ICT in classroom teaching. By presenting an historical overview of the usage of pen-based input devices in classrooms and lecture halls and two case studies with currently available technology, we demonstrated that today’s hardware equipment can be used to greatly improve the teaching experience. Using them, we are able to build a ubiquitous computing environ-
Human Computer Interaction for Computer-Based Classroom Teaching

ment in the classroom that offers teachers full interaction with the respective technology using a pen while at the same time not interfering with the interaction between teacher and students.

As a trend of moving toward natural environments in the digital world, such a setup encourages freehand writings to become the bridging tool that encompasses portable and synchronizable standards of manipulating ink data between mediums and documents. Hence, we introduced the digital ink as a new “first-class” data type serving two purposes, first, the representation and preservation of information through freehand writing and annotations on the board. In this context we discussed issues about the personalization of digital ink as well as its archival, later access and streaming replay. Subsequently, we described the second purpose of digital ink—the operation and control of the classroom’s ICT via pen-gestures. By interpreting ink traces as gestures, users can interact with the ubiquitous computing environment in a more interactive and intuitive way than with regular interface designs using traditional graphical user interfaces developed and optimized for desktop computing. This has also been proven to be the case for multi-user interaction on the horizontal table-interface: (cf. Figure 8) recognizing a gesture in its respective interpretation also tells us the side of the table it originated from, thus simultaneously identifying the user seated at that corner without having to install additional sensing devices.

For the future, we expect further improvements in the respective hardware and the development of new devices that might further improve the usage of ICT in the classroom. Although current devices and hardware configurations provide a real alternative to the traditional blackboard and overhead projector equipment, as we described in the hardware-related part of this chapter, there is always room for improvement in terms of size, quality, and production costs. At the same time, new technology keeps evolving. Interesting developments with great potential for classroom usage include, for example, the so called electronic paper/electronic ink approaches (see Jacobson, Comiskey, Anderson, & Hasan, 2005 and Xerox Parc, 2005).

Wireless networks are more and more installed at universities, and although not directly related to input devices, this infrastructure will have great influence on the ICT usage in classrooms and thus on the HCI. For example, it is not clear how to seamlessly integrate different devices with different sizes into the overall communication and learning process. Lecturers write on large electronic whiteboards, while students connect mobile devices, such as laptops, Tablet PCs, or even very small PDAs and SmartPhones, to the wireless LAN. How to represent and exchange information between such devices that differ in size and processing power is an open research problem (MacIntyre & Feiner, 1996) where convincing solutions are still missing.

In addition, there exists an obvious trend toward computer-supported group work in classrooms. Instead of purely instructional presentations and lectures, which have been the focus of this chapter, other ways of teaching, such as seminars and group work, can benefit from the usage of ICT as well. Important research in this area has been done, for example, by the Interactive Workspaces Project at Stanford University (Johanson, Fox, & Winograd, 2002). In our own research lab, we are currently exploiting the usage of an “interactive table” where a data projector is mounted on the ceiling, projecting the content of a computer screen onto a horizontally mounted electronic whiteboard (cp. Figure 8). While the respective hardware to build such a table is already available, many open questions and problems remain regarding the design of the software side of the user interface, such as dealing with different orientations when people sit on different sides of the table, enabling people to reach remote objects on the virtual table’s surface (i.e. the projected computer screen), and so forth. Such horizontally mounted interactive computing systems are an ac-
Human Computer Interaction for Computer-Based Classroom Teaching

Figure 8. An electronic whiteboard mounted horizontally in order to create an ‘interactive table’ for group-work

NOTE

Some of the notions for tools and devices described in this article are registered Trademarks of the respective companies or organizations. We kindly ask the reader to refer to the given references. All features and characteristics of these systems have been described to the best of our knowledge. However, we do like to mention that specific characteristics and technical specification frequently change, and therefore discrepancies between the descriptions in this article and the actual systems are possible.

REFERENCES


This work was previously published in Enhancing Learning Through Human Computer Interaction, edited by E. McKay, pp. 21-42, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 4.3
Wikis as an Exemplary Model of Open Source Learning

Robert Fitzgerald
University of Canberra, Australia

ABSTRACT

In their simplest form, Wikis are Web pages that allow people to collaboratively create and edit documents online. Key principles of simplicity, robustness, and accessibility underlie the wiki publication system. It is the open and free spirit of Wikis fundamental to open source software (OSS) that offers new contexts for learning and knowledge creation with technology. This chapter will briefly consider the role of technology in learning before discussing Wikis and their development. The emerging literature on the application of Wikis to education will be reviewed and discussed. It will be argued that Wikis embody an exemplary model of open source learning that has the potential to transform the use of information communication technologies in education.

INTRODUCTION

Wikis are an instance of what is known as a read/write technology. They allow groups of users, many of whom are anonymous, to create, view, and edit Web pages. In many cases, these pages are online, but there are instances of Wikis used as personal notebooks (e.g., Tiddlywiki, http://www.tiddlywiki.com/). All wiki systems use a simplified html markup language, but as their use spreads, so does the appeal of more user-friendly java-based WYSIWYG editors. It will be argued that the simplicity, accessibility, and openness of wikis support a model of collaboration and knowledge building that represents an exemplary model of learning with technology. This model is congruent with many of the key principles embodied in free and open source software (FOSS) and sociocultural theories of learning (Lave & Wenger, 1991; Vygotsky, 1978; Wenger, 1998). Many Internet-based communities and groups are already finding ways to embrace these forms of learning as a part of their ongoing process of community capacity building. In contrast, formal places of learning such as schools and universities have been slow to explore the potential of this technology. This chapter will briefly consider the role of technology in learning before discussing Wikis and their development. The chapter argues that FOSS and Wikis in particular offer education
far more than just low-cost software or even sound principles of practice; they open up a space for new models of learning and knowledge creation with technology. The emerging literature on the application of Wikis to education will be reviewed before considering Wikis as an exemplary model of open source learning.

TECHNOLOGY AND LEARNING

While the application of computing technology to teaching and learning has at least a 30-year history, there is a large body of literature that suggests education is still struggling to use technology effectively (Cuban, 2001; Healy, 1998; Oppenheimer, 2003; Postman, 1993; Stoll, 1999). Results from large international studies show that the dominant use of technology tends to focus on skills and involves learners as content users and not content creators (Kozma, 2003). Part of the problem is that formal places of learning by their very nature are highly structured contexts. The role of context is important because there is a direct relationship between form and quality of the pedagogy and the form and quality of the resultant learning. Different teaching approaches and learning contexts result in different outcomes for students. In Boaler’s (1997) study of mathematics classrooms, she showed that teacher-centered and rule-based teaching approaches not only produce low levels of student engagement but work to effectively limit the scope of the learning outcomes. There is a strong suggestion from her work that routine-style classrooms generate routine knowledge and that this is neither of the quality nor quantity required for real-world mathematical problem solving. Her key finding is that context matters. The Russian neuropsychologist Alexandra Luria understood this relationship well when he argued that cognition is a function of context. “Cognitive processes … are not independent and unchanging ‘abilities’ … they are processes occurring in concrete, practical activities and are formed within the limits of this activity” (Luria, 1971, p. 266). In effect, Luria was saying that cognition is plastic, a finding that has subsequently been confirmed by contemporary neuroscience (Goldberg, 1990, 2001). The activities and tasks we set for learners not only determine the type and quality of knowledge that is produced but, more importantly, set the parameters for the development of their cognitive processes. Therefore, from a philosophical and practical design point of view, the contexts or settings of learning should be as open and free as possible.

There is also a growing body of literature suggesting that young people learn in different ways to past students and, therefore, require (and even demand) different teaching approaches (Gee, 2003; Oblinger, 2004; Prensky, 2001). Chris Dede (2005) has written extensively in this area, and in his assessment, these learners seek to co-design their learning experiences and prefer communal learning over individual learning. Anyone who has recently studied in schools or universities will know that despite institutional rhetoric to the contrary, these new modes of teaching and learning are not widespread. The pedagogical challenge is to use technology in ways that build upon learners’ existing experiences and foster the creation of what von Krogh, Ichijo, and Nonaka (2000) refer to as communities of knowledge. In education, there is widespread recognition of the need to explore more collaborative approaches to learning (Jonassen, Peck & Wilson, 1999; Kozma, 2003; Laurillard, 2002; Sefton-Green, 2004; Somekh, 2004). Wikis offer one such tool, which is already a part of many learners’ everyday lives as are a wide variety of other social software such as blogs and social networking applications like MySpace (http://www.myspace.com/).

BACKGROUND TO WIKIS

The founding developer of the World Wide Web (WWW), Sir Tim Berners Lee, first conceived of
the Internet as a way for people to both read and write data. The reality of nonintuitive operating systems, html-based coding, clunky file transfer tools, and security restrictions guaranteed that while we could all read the Web, very few of us were able to easily write and publish material. To address some of these issues, Ward Cunningham developed the first Web-based Wiki (http://en.wikipedia.org/wiki/WikiWikiWeb) in 1995 as part of the Portland Pattern Repository (PPR). The origin of wikis, however, goes back much earlier to the 1980s and Cunningham’s work on hypertext using Hypercard™ (eWEEK, 2006). In his original formulation (see Table 1), Cunningham developed wiki software to adhere to a number of core design principles (http://c2.com/cgi/wiki?WikiDesignPrinciples).

These basic principles result in a Web publication system that is simple, robust, and accessible. The system is designed to position users differently from traditional publishing models—from a Web reader to a Web author. This can be seen particularly in the universal principle in which any writer is automatically both a space organizer and an editor.

**FREE AND OPEN LEARNING**

Free and open are key principles underpinning FOSS. This chapter argues that these also underpin learning with Wiki technology. For a piece of software to be open source, it must also adhere to four freedoms summarized in Table 2 (http://www.gnu.org/philosophy/free-sw.html).

While these freedoms relate to OSS, they are enacted within a community of software developers in which community learning and knowledge construction is central. In this sense, open can be used to refer to something that is visible and without barriers. In relation to FOSS, it also refers to a principle of practice that embodies an attitude of generosity. This is perhaps best thought about in terms of an individual (or group) that makes himself or herself available to others and is free and willing to think about new ideas. The notions of free and open have much to offer our conceptions of learning and the underpinning processes of collaboration and knowledge construction with Wikis.

**Table 1. Wiki core design principles**

<table>
<thead>
<tr>
<th>Design Principle</th>
<th>Core functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>Should a page be found to be incomplete or poorly organized, any reader can edit it as they see fit</td>
</tr>
<tr>
<td>Incremental</td>
<td>Pages can cite other pages, including pages that have not been written yet</td>
</tr>
<tr>
<td>Organic</td>
<td>The structure and text content of the site are open to editing and evolution</td>
</tr>
<tr>
<td>Mundane</td>
<td>A small number of (irregular) text conventions will provide access to the most useful page markup</td>
</tr>
<tr>
<td>Universal</td>
<td>The mechanisms of editing and organizing are the same as those of writing so that any writer is automatically an editor and organizer</td>
</tr>
<tr>
<td>Overt</td>
<td>The formatted (and printed) output will suggest the input required to reproduce it</td>
</tr>
<tr>
<td>Unified</td>
<td>Page names will be drawn from a flat space, so that no additional context is required to interpret them</td>
</tr>
<tr>
<td>Precise</td>
<td>Pages will be titled with sufficient precision to avoid most name clashes, typically by forming noun phrases</td>
</tr>
<tr>
<td>Tolerant</td>
<td>Interpretable (even if undesirable) behaviour is preferred to error messages</td>
</tr>
<tr>
<td>Observable</td>
<td>Activity within the site can be watched and reviewed by any other visitor to the site</td>
</tr>
<tr>
<td>Convergent</td>
<td>Duplication can be discouraged or removed by finding and citing similar or related content</td>
</tr>
</tbody>
</table>
EXAMPLES OF WIKI PROJECTS

The use of Wikis had a significant development when Jim Wales launched Wikipedia (http://en.wikipedia.org) and the Wikimedia Foundation (http://wikimediafoundation.org) in 2001. For the first time, this project made Wiki technology and tools freely and readily available to the public while also offering a series of projects to which they could contribute. The most recognized Wiki is the free-content encyclopedia Wikipedia (http://en.wikipedia.org/wiki/), which is maintained by a group of volunteers from around the world and is generally regarded as a rich information resource (LeLoup & Ponerio, 2006; Lih, 2004).

Jim Wales explained:

Wikimedia’s mission is to give the world’s knowledge to every single person on the planet in their own language. As part of that mission, Wikipedia is first and foremost an effort to create and distribute a free encyclopedia of the highest possible quality. Asking whether the community comes before or after this goal is really asking the wrong question: the entire purpose of the community is this goal. (Wales, 2005)

Wikipedia currently comprises 1,315,437 English language articles and nearly two million registered accounts (http://en.wikipedia.org/wiki/Special:Statistics). Roughly 25% of these articles have been translated into other languages (nine at last count). Since July 2002, Wikipedians have made more than 70 million edits. Web traffic statistics rank Wikipedia as the 17th most popular Web site (out of 250 million) on the Internet with more than 2.9 billion page views in the month of August 2006 (http://www.alexa.com). Currently there is a number of related Wikimedia Foundation projects that are in development, including Wikipedia (http://www.wikipedia.org/); Wiktionary (http://wiktionary.org/); Wikibooks; Wikinews (http://www.wikinews.org/); Wikiquote (http://www/wikiquote.org/); Wikicommmons (http://commons.wikimedia.org); Wikisource (http://wikisource.org/); Wikispecies (http://species.wikimedia.org) and the soon to be officially established Wikiversity (http://en.wikibooks.org/wiki/Wikiversity). Each of these projects takes the basic wiki model and extends it into a specific area with a specific goal. For example, Wikibooks (http://en.wikibooks.org) is a project designed to produce open-content textbook modules to create global curricula. To date 21,019 book modules have been developed for more than 1,000 books. Some of these books are available in PDF format.

The large-scale application of wikis to community knowledge building can also be seen in Wikia (http://www.wikia.com) and Wikispaces (http://www.wikispaces.com/). The aim of these initiatives is to provide individuals and communities with a Web site they can use to create open content around their areas of interest. For example, establishing a Wikia site requires the topic to appeal to a large number of people and

<table>
<thead>
<tr>
<th>Freedom</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>The freedom to run the program, for any purpose</td>
</tr>
<tr>
<td>1</td>
<td>The freedom to study how the program works, and adapt it to your needs Access to the source code is a precondition for this</td>
</tr>
<tr>
<td>2</td>
<td>The freedom to redistribute copies so you can help your neighbour</td>
</tr>
<tr>
<td>3</td>
<td>The freedom to improve the program, and release your improvements to the public, so that the whole community benefits. Access to the source code is a precondition for this</td>
</tr>
</tbody>
</table>
that its content will have some longevity. Computer game players have been particularly active, creating communities around their games. One example of how this is used can be seen in the ways that players of the massively multiplayer online game Runescape (http://www.runescape.com/) have built encyclopedic knowledge about all aspects of the game (http://www.wikia.com/wiki/Runescape).

**LITERATURE REVIEW ON WIKIS**

Many of the applications of Wikis are entirely congruent with the so-called Web2.0 and social software models that attempt to offer simple and robust technologies to non-expert users so they are able to create content and build communities. The last few years have seen a proliferation of simple tools for authoring Web content, particularly in the area of personal blogging. Wikis are still without peer when it comes to the large-scale collaborative authoring of Web content (Lamb, 2004; Wagner, 2004). A literature based on the application of Wikis is beginning to emerge, although to date it is primarily conceptual and descriptive. Generally, there is widespread agreement that Wikis represent an innovative and potentially powerful tool for collaborative content creation and sharing (Lamb, 2004; Wagner, 2004). There have been few studies that have tested empirical propositions, although many of these have been exploratory in nature.

Wikis have been studied in language learning (Godwin-Jones, 2003; LeLoup & Ponerio, 2006; Wang et al., 2005; Wei et al., 2005) as tools in higher education (Augar, Raitman, & Zhou, 2004; Bold, 2006) to promote forms of participatory journalism (Lih, 2004), as a tool for story-telling in primary schools (Désilets & Paquet, 2005), and examined for their potential role in increasing citizen participation in e-government initiatives (Wagner, Cheung, & Ip, 2006).

In summary, the empirical work finds the following:

- The effective use of Wikis appears dependent on a clear goal matched to a group of committed uses (Godwin-Jones, 2003).
- Highly structured environments that rely on top-down approaches (as opposed to bottom-up) limit the potential of Wikis as a tool for learning (Engstrom & Jewett, 2005; Wagner, 2004).
- Wikis such as Wikipedia are a rich source of information that can promote content creation, sharing, and discussion (LeLoup & Ponerio, 2006, Lih, 2004).
- It is important to augment students, Wiki work with strategies to promote deep and critical thinking to ensure high quality work emerges (Engstrom & Jewett, 2005).
- Wikis support a short edit-review cycle that ensures the rapid development of content (Lih, 2004).
- Employing the user as organizer and editor (many “eyeballs”) is a highly effective strategy for ensuring quality (Lih, 2004).

There have been widespread calls for more research on Wikis (Lih, 2004; Wagner, 2004; Wei et al., 2005). It will be particularly important to develop research methods that are sensitive to both the quality of content produced in concert with how this content emerges within a community of learners. There is encouraging work on the development of metrics to assess the quality of Wiki (see Wikipedia) articles based on edit histories (Lih, 2004), but we also need to examine and assess the quality of the articles. Wiki edits are easily quantified, but what they relate to is not always clear. For example, a recent study found an inverse relationship between the quantity of Wiki edits and final exam scores (Wang et al., 2005). The authors advised caution against interpreting these findings as evidence that wikis are counter-productive to learning, but it does highlight the
need for more nuanced and in-depth empirical studies on Wikis.

Wikis have generated considerable interest in education because they appear to support more collaborative models of teaching and learning. It is fair to say that there is considerable anecdotal evidence that Wikis can and should play a key role in e-learning in support of a more conversational and dialogic approach to knowledge creation and sharing.

AN EXEMPLARY MODEL OF OPEN SOURCE LEARNING

Wikis offer a different model of creating, editing, and sharing knowledge that is consistent with the educational push toward what have become known as sociocultural or constructivist approaches to learning. A founding thinker in these areas, Lev Vygotsky (1978), contended that learners neither receive knowledge nor simply discover it. They learn in social contexts in interaction with both humans and tools. A key concept for Vygotsky was the zone of proximal development (ZPD) in which he said all learning takes place. In Vygotsky’s basic model, it is adults who scaffold young learners, helping to extend their thinking and learning. However, as the technological tools develop and evolve, we are beginning to see ways that both humans and their tools can scaffold learning. The technological spaces that make up wikis enable new forms of sociotechnological ZPDs that support both individual and community knowledge creation.

This focus on community and the power of joint construction is taken up in The Wisdom of Crowds (Surowiecki, 2004). Surowiecki argues that the collective knowledge of large groups is often unrecognized and almost always undervalued by society. He explains that many everyday activities, from voting in elections and the operation of the stock market to the way Google locates Web pages, depend on the collective input and knowledge of large groups. Of course, not all crowds are smart, but Surowiecki believes that under the right conditions, crowds can act more wisely than an expert individual. To achieve the best results, crowds must be able to support diversity of opinion, relative independence in an individual’s thinking, a model of decentralization that allows individuals to draw on their local knowledge and aggregation; that is, embody a process whereby individual knowledge can be combined into an integrative whole. The ways technology might be used to support the development of smart crowds is a relatively unexplored area; however, applications such as Wikis, blogs, and multiplayer games certainly show how large groups of people can productively interact online. While there is a huge qualitative difference between group interaction and wisdom, there is a relationship to be explored that highlights the importance of developing large-scale social technologies such as Wikis.

The previous discussion suggested that collaborative knowledge creation should be an important feature of formal learning; however, to date, and particularly with reference to the use of technology, it has not been. Some of the uses of technology by young people in their everyday lives seem to get closer to this goal. The way they work in their communities around multiplayer games to talk to each other and build knowledge is one example. The next section considers the potential that Wikis offer to achieve this goal.

One of the strengths of Wiki software is its capability to document and record aspects of the knowledge creation process. From an educational point of view, this can provide valuable insights to the knowledge construction process. In most Wikis, an article features a number of views: the article page, a discussion page, article editor, and history. A rich edit history features full revision history permitting comparison between current and last entry. Edits can further be identified via flags and commenting, helping others understand the changes that have been made. To help assure quality, edits also appear on a recent-changes page.
Rules can be set up to show pages that have been changed since the last visit. A list of contributions by users offers various analyses, often with full history and comparison tools. In combination, these tools open up the possibility of exploring the relationship (and tension) between individual and group constructions.

Surowiecki said:

*Any “crowd”—whether it be a market, a corporation, or an intelligence agency—needs to find the right balance between the two imperatives: making individual knowledge globally and collectively useful (as we know it can be), while still allowing it to remain resolutely specific and local. (Surowiecki, 2004, p. 140)*

Wikis allow both individual contributions and the evolving group product to sit alongside each other. The examples of Wikis outlined here seem able to create new forms of sociotechnological ZPDs for learners. These zones support both individual and community knowledge creation in ways that are consistent with the notion of communities of practice (Lave & Wenger, 1991; Wenger, 1998).

Wikis appear well suited to building knowledge in which the representation of balanced opinion is valued. While there is no guarantee that this prevents wikiwars, it does seem that strong opinion is better suited to other spaces such as blogs. It is clear from the work of the Wikimedia Foundation that there are no hard and fast rules to using Wikis. What is apparent, though, is that many of the more successful projects embody a spirit of community characterized by openness and freedom.

**CONCLUSION**

In the end, the success of innovations in learning such as Wikis will be seen in the increased capacity of individuals and their communities to create and apply new knowledge. Incorporating tools that not only facilitate but also document the effective management of information and creation of knowledge is now essential for an innovative and productive 21st-century society. Wikis are significant and innovative because they attempt to position learners as knowledge creators rather than simply content users. They also represent the application of new collaborative technologies in ways that are free and open. In terms of education, the Wiki model locates the challenge of improving information, data management, and knowledge creation processes within a community model. It also builds on strategies for increasing the capacity of all community members to expand their ways of thinking creatively and working collaboratively.

A key feature of this approach is that knowledge and practice are shared in a spirit of generosity. The extensive use of Wikis in education begs the question: Are we really prepared to engage in this form of open source learning?

**REFERENCES**


Wagner, C. (2004). Wiki: A technology for con-
versational knowledge management and group collaboration. *Communications of the Association for Information Systems*, 13, 265-289.


**KEY TERMS**

**Constructivist:** An approach based on the work of Lev Vygotsky, who contended that learners neither receive knowledge nor simply discover it. They learn in social contexts in interaction with both humans and tools.

**Free and Open Source Software (FOSS):** A term first described by Richard Stallman referring to a software development process in which the software source code is made freely available for subsequent modification and development.

**Hypercard:** A hypermedia program developed by Apple Computer in the 1980s.

**Open Source Learning:** A model of learning inspired by the key principles or freedoms embodied in the FOSS movement.

**Web2.0:** A term coined by Tim O’Reilly (http://tim.oreilly.com/) referring to a range of second-generation Web publishing and social networking technologies.

**Wiki:** A form of read/write technology that allows groups of users, many of whom are anonymous, to create, view, and edit Web pages.

**Wikia:** A project to provide communities with Wiki-type Web sites (see http://www.wikia.com).

**Wikimedia Foundation:** An international nonprofit organization run by Jim Wales, using wiki technology to promote free and open large-scale collaborative content creation projects (http://wikimediafoundation.org).

**Wikipedia:** A free-content encyclopedia (http://en.wikipedia.org/wiki/).

**Zone of Proximal Development (ZPD):** The difference between what learners can do by themselves and with the assistance of more capable adults or peers.

ABSTRACT

Over the past 15 years a considerable amount of research has been devoted to study of the socioeconomic aspects that affect the use of technology in the mathematics classroom. With the call for curricular and instructional reform, educational institutions have embarked on the process to reform their educational practices to aid the urban student in their quest to obtain quality mathematics and science based education with the integration of technology. The study performed was to reexamine the socioeconomic disparities in the mathematics classroom as it relates to implementing technology interventions to support quality teaching and active student learning. This article is to provide empirical evidence of whether these disparities continue to exist and their effects on student achievement in the mathematics classroom. The results of this study showed an overall positive relationship regarding the use of technology interventions within the mathematics classroom with levels of student achievement,
showing a clear signs of continued disparities within mathematics classroom.

INTRODUCTION

The introduction of microcomputers into classrooms during the 1980's was heralded by many as the dawn of a new era in American education. Proponents argued that technology had the potential to fundamentally transform the nature of teaching and learning (Papert, 1980; U.S. Congress, Office of Technology Assessment, 1988). However, over time, it has become apparent that it is far easier to acquire hardware, software, and Internet access (Becker, 1991; Dividing Lines, 2001) than it is to capture the potential of technology in significantly meaningful outcomes (Cuban, 2001). Likewise, educators concerned about the chronic underachievement of students often fall prey to the allure of technology as a tool for reversing the historical influences of poverty, discrimination, inequity, chronic underachievement, and lack of opportunity. However, 25 years after the introduction of the computer into the classroom, many of the expectations associated with technology in education remain unrealized to some but to other, technology has proven to be an effective tool in the efforts to provide students with opportunities for quality teaching and active student learning and engagement.

Educational institutions have called for instructional and curriculum reform that includes active engagement of students, quality assessments, and the increased and innovative use of technology applications to promote quality teaching and active student learning (U.S. Department of Education, 2001). This is true in the field of mathematics where organizations such as the National Council of Teachers of Mathematics (1989, 2000), Mathematical Science Board (MSEB, 1991), and Mathematics Association of America (1991) have stressed that technology is essential in teaching and learning mathematics. The underlying assumption of these organizations and math educators alike is that technology will enable students to explore mathematics with a greater depth and will allow them to study topics that were previously impractical (Garofalo, Drier, Harper, Timmerman, & Shockey, 2000). However, in order for technology to have greatest impact on our educational system, all students must have access to technology. For that reason technology has the potential to narrow the achievement gap, if equally distributed or widen the gap if only accessible to selected groups in the educational system (Kulik, 2002; Waxman, Connell, & Gray, 2002).

REVIEW OF LITERATURE

Over the past 15 years a considerable amount of research has been devoted to sociocultural disparity in technology availability and use in the mathematics (Becker, 2000; Garofalo et al., 2000; Means et al., 2001; Manoucherhri, 1999; National Center for Educational Statistics, 2004; Owens, 1993; Owens & Waxman, 1994, 1995; U.S. Department of Education, 2001; Huang & Waxman, 1996). Past studies conducted by Becker (2001) and Coley, Cradler, and Engel (1997) found students from higher income families have been found to use computers in school and in their homes more frequently than students from lower-income families. Students of color from urban schools have also been found to have less access to computers compared to Anglo-suburban students (Owens & Waxman, 1993, 1994). More recently, lower SES schools are only half as likely to have high speed Internet compared to high SES schools (Advanced Telecommunications, 1997). Consistent with this idea of access are the issues within the digital divide itself. Within the past decade, a growing body of evidence supports the ever-widening technological gap among members of society, in particular children and the elderly (NTIA, 1995, 1997, 1999), with a important emphasis on urban...
schools with the inner cities. The groups identified who lack access to information and technological resources include people of color, specifically African American and Hispanic Americans, those who are poor and of the working class, individuals of low-income, those who possess less than a high school level of education, children of single parents, and residents of inner-cities (NTIA, 1995, 1997, 1999, 2002, 2004). Access is key in this issue.

Despite the constraints on school funding in most states, schools have devoted an increasing percentage of their annual budgets to technology. The majority of the efforts of the educational community over the past decade to acquire hardware, software, and Internet access have been successful (Dividing Lines, 2001; Education Week, 2004). However, clear evidence of a digital divide, parallel to historical disparities, continues to distinguish urban schools from their affluent counterparts (Chen & Thielemann, 2001; Guttentag & Eilers, 2004; National Center for Education Statistics, 2004). Historical measures of digital equity have been based on the ratio of the number of computers divided by the number of students. A more recent measure involves determining levels of Internet access. Another dimension of this problem relates to questions about differences in home access to technology, therefore impacting urban student achievement.

A second source of disparity in technology use is how technology is used. Previous studies conducted by Becker (2001) and Finneran (2000) found that low SES schools are more likely to use technology for drill and practice, whereas high SES school uses technology in innovative teaching strategies. This idea is consistent with the ideas of curriculum reform and reconceptualization put forth by Pinar (2002). Pinar (2002) suggests that the instruments of computer technology are used to drill and kill students into passing standardized test, not actually being integrated effectively into classroom instruction or pedagogical practice that promote quality teaching and active student learning. He further explains that the current use of computer technology in urban schools generally serve to turn its users (students) into disembodied and alienated subjects. Furthermore, as explained in Becker (2001) high SES students are more likely to use technology for school assignment, use e-mails, and use educational programs.

Finally, educational reform programs such as the No Child Left Behind Act of 2001 has focused on accountability and high-stakes testing. The renewed importance of test scores will fundamentally alter the attention and activities of low-achieving schools as they seek to achieve the required benchmark scores (Johnson, 2002). As a result, it was anticipated that previous technology buying sprees will be redirected. One approach to this redirection is that there is new emphasis placed on technology initiatives that enhance schools’ assessment and early intervention efforts (Education Week, 2003) that feature learning activities with instructional objectives and record keeping.

A third source of disparity in technology use deals with the nature of technology adoption and organizational change. A thorough analysis of major research related to technology and teacher motivation, adoption and usage uncovered important factors that are involved in determining their willingness to use such approaches in the teaching and learning process (Braak, 2001; Chiero, 1997; Fabry & Higgs, 1997; Solomon & Wiederhorn, 2000). The literature indicates that teacher’s willingness to adopt and implement learning technologies for the teaching and learning process varies, but all share the same denominator—proper internal and external motivation (Braak, 2001; Chiero, 1997; Fabry & Higgs, 1997; Mooji & Smoot, 2001; Solomon & Wiederhorn, 2000). The top reasons teachers choose to use technology for the teaching and learning process dealt with the notion of continuous training and development, proper technology support from technology personnel, encouragement from school administration, and an organizational structure that sup-
ported teachers using technology (Grank, Zhao, & Borman, 2004). While these are the key issues ascertained from the research study, the ideas of change management, technology adoption and innovation, self-efficacy and motivation, support, and the computing experience are relevant to this discussion. The American Council of Education (1999) report noted that the quest to optimize the use of technology in schools could not be fulfilled by the mere supply of more hardware and software. The report suggested that what needs to be addressed first are ways identify, motivate, and equip teachers with the skills necessary for effective use of educational related multimedia technologies in their classroom practices.

In contrast to the indicators that profile the acquisition of technology, less information is typically available to reflect progress toward implementing technology applications that enhance the teaching and learning process (NTIA, 1995, 1997, 1999, 2000, 2002; CEO Forum, 1997). It is rather a task of ease to promote and profile computer-to-student ratios in school reform experiments with little regard to how technology is used in the urban classroom (Brainbridge, Lasley, & Sundre, 2004; Guttentag & Eilers, 2004). The use of technology in the classroom has moved through definable periods involving programming, computer-assisted instruction, problem-solving environments, personal productivity, Web-based instruction, and hypermedia. Problems associated with limited opportunities for teacher professional development to learn about new innovations (Lonergan, 2001), as well as limited funds for new hardware and software, often result in the routine use of student learning activities that have been abandoned by high-performing schools (Burnett, 1994). As a result, it is essential that questions be raised about the nature of technology learning activities in urban schools when expectations are held that technology is used to enhance academic achievement. While looking into the use of technology in the urban school environment, it has been noted that the use of these tools does not reflect innovative uses in integration into the curriculum for active student learning, engagement or to enhance the teaching process. Typically urban schools have antiquated resources and in adequate technological support for the use of technology in classroom.

Although these studies have established a pattern for technology disparities in the past, most of these studies rely on data collected in the mid to late 1990’s. Moreover, the last national report on the status of technology use entitled “Teachers’ Tools for the 21st Century: A Report on Teachers’ Use of Technology” was published in 2000, however, it used data from the 1999 FRSS survey (National Center for Educational Statistics, 1999). Therefore, updated studies are needed that examine current data to determine where we are in the quest to narrow the achievement gap with the aid of technology. The purpose of the present study is to reexamine technology use and to provide evidence of whether or not disparity issues still exist using the latest national survey produced by the National Center for Educational Statistics (2002). Our focus in this report is on the technology practices in the mathematics classroom.

METHODS

The data for the present study was drawn from the base year of the Educational Longitudinal Survey of 2002-2004 (ELS: 02). The ELS:02 is a national survey administered by National Center for Educational Statistics to provide an overall picture of the present educational experience of high school students in the U.S. The survey has four subcomponents completed by students, teachers, parents, and school administrators. The design used to collect the data was a two-stage, stratified national probability sample. The survey included 1,052 public and private school representing about 12,000 students across the country. For the present study, only the student survey data was used.
The student survey provided items used to measure how often technology is being used in mathematics classrooms along with how this technology is being used as reported by students. The items used to measure how often technology is being used in the classroom were the following: (a) how often do you use calculators in your mathematics classroom?; (b) how often do you use graphing calculators in your mathematics classroom?; and (c) how often do you use computers in your mathematics classroom?. The outcome measures for these items were a five-point Likert scale, ranging from “never” to “everyday or almost.”

To measure how technology is being used in mathematics classrooms the following items were used: (a) how often do you use a computer to review your math work in your mathematics classroom?; (b) how often do you use a computer to solve math problems in your mathematics classroom?; (c) how often do you use a computer for graphing of math problems in your mathematics classroom?; (d) how often do you use a computer to practice math drills in your mathematics classroom?; (e) how often do you use a computer to analyze data in your mathematics classroom?; (f) how often do you use a computer to apply learning in your mathematics classroom?; (g) how often does your math teacher use a computer for one-on-one instructions in your mathematics classroom?; and (h) how often does your math teacher use computer to show new topics in your mathematics classroom?. The outcome measures for these items were also a five-point Likert scale, ranging from “never” to “everyday or almost”. The survey also provided a measure of student’s socioeconomic status. This measure was developed as a composite variable constructed from five equally weighted, standardized components: father’s education, mother’s education, family income, father’s occupation, and mother’s occupation. The distribution of these SES values were divided into four quartiles, ranging from lowest to highest. To determine if an association exists between technology use and students’ socioeconomic status a Pearson’s chi-square was used.

RESULTS

Table 1 reports the results of the frequency of calculator and computer use in mathematics classrooms. The results indicated that students are using more calculators in their math classrooms compared to computers. 58% of the students reported that they had used calculators every day in their math classroom compared to 8% that indicated they were using computers on a daily basis. 30% of the students reported using the graphing calculator on a daily basis. One-third of these students reported they never use the graphing calculator in their classroom. 61% of the students indicated they never use computers in their math classroom. Finally, 7.4% indicated they were using computers on a daily basis.

Table 2, compares calculator and computer use across socioeconomic levels. The results indicate that a significant association (p<.001) exists between calculator use and socioeconomic levels. In this case, the lowest SES group reported using calculators on a daily basis the least. On the other hand students in the highest SES group reported using calculators on a daily basis more often. 48% of the lowest SES group reported using calculators compared to about twice as many students (68%) classified in the highest SES group. There was also a significant association (p < .001) between daily use of graphing calculators and SES group membership. 21% of the students classified in the lowest SES reported using the graphing calculator on a daily basis compared to about twice as many students (42%) classified in the highest SES group. The final comparison looked at computer usage across SES levels. The results also indicated a significant association exists (p < .001) between computer usage and SES levels. In this case, students classified in the lowest SES group were more likely to use computers computer compared to those students.
Re-Examining the Socioeconomic Factors Affecting Technology Use in Mathematics Classroom

Table 1. How often technology is being used in mathematics classrooms overall frequency

| N=11,618 | How often do you use calculators in your math class | 6.2 | Never |
| | | 12.0 | Rarely |
| | | 7.7 | Less than once a week |
| | | 18.0 | Once or twice a week |
| | | 58.0 | Everyday or almost |
| | How often do you use graphing calculator in your math class | 33.1 | Never |
| | | 19.6 | Rarely |
| | | 3.3 | Less than once a week |
| | | 2.2 | Once or twice a week |
| | | 8.8 | Everyday or almost |
| | How often do you use computers in your math class | 60.7 | Never |
| | | 2.2 | Rarely |
| | | 7.7 | Less than once a week |
| | | 7.8 | Once or twice a week |
| | | 7.4 | Everyday or almost |

Chi. Sq = 341.23, p < .001***

Table 2. Socioeconomic status and how often technology is being used in mathematics classrooms percentage of use

| N= 11,618 | How often do you use calculators in your math class | 8.83% | 6.67 | 5.91 | 4.12 | Never |
| | | 18.12 | 12.91 | 11.08 | 7.18 | Rarely |
| | | 6.78 | 6.24 | 5.24 | 4.82 | Less than once a week |
| | | 19.57 | 18.94 | 17.68 | 16.24 | Once or twice a week |
| | | 46.71 | 55.25 | 60.08 | 67.63 | Everyday or almost |

Chi. Sq = 341.23, p < .001***

| How often do you use graphing calculator in your math class | 37.65 | 37.30 | 32.98 | 25.83 | Never |
| | | 22.95 | 20.47 | 20.30 | 15.66 | Rarely |
| | | 16.16 | 6.24 | 6.46 | 5.56 | Less than once a week |
| | | 31.31 | 10.96 | 11.15 | 11.42 | Once or twice a week |
| | | 94.94 | 25.00 | 29.11 | 41.53 | Everyday or almost |

continued on following page
Re-Examining the Socioeconomic Factors Affecting Technology Use in Mathematics Classroom

Table 2. continued

<table>
<thead>
<tr>
<th>How often do you use computers in your math class</th>
<th>53.73</th>
<th>61.06</th>
<th>62.81</th>
<th>64.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never</td>
<td>21.74</td>
<td>19.50</td>
<td>18.85</td>
<td>20.92</td>
</tr>
<tr>
<td>Rarely</td>
<td>13.13</td>
<td>5.00</td>
<td>6.00</td>
<td>5.80</td>
</tr>
<tr>
<td>Less than once a week</td>
<td>53.53</td>
<td>6.24</td>
<td>5.28</td>
<td>4.09</td>
</tr>
<tr>
<td>Once or twice a week</td>
<td>9.86</td>
<td>8.19</td>
<td>7.04</td>
<td>5.10</td>
</tr>
</tbody>
</table>

Chi. Sq = 375.88, p < .001***

Table 4 compares computer use across socioeconomic levels. The results indicated that the higher computer use was significantly associated with low SES classification. Moreover, as SES Levels increased computer usage decreased. The most frequent use of computers was to solve math problems. The results indicate that about 21% of the low SES students compared to about 10% of the high SES students reported they used computers to solve math problems everyday or almost everyday. Students classified as second and third SES levels reported about 18% and 15% of the times they used computers to solve math problems everyday or almost everyday respectively. The next most frequent use of computers was to apply learning. Students classified as low SES indicated that used computers for this purposes everyday or almost everyday about 18% and 15% of the times they used computers to solve math problems everyday or almost everyday respectively. The next most frequent use of computers was to apply learning. Students classified as low SES indicated that used computers for this purposes everyday or almost everyday about 17% of the times compare to about 9% for the high SES students. The second and third level SES students indicated they used computers everyday or almost everyday to apply learning about 15% and 13% of the times respectively. Similarly, about 17% of the lowest SES group reported using computers to practice math drills compared to about 10% of the high SES students everyday or almost everyday. Furthermore, about 30% of the students reported that they never used the computer for graphing math problems.

Table 3 reports the frequency of how computers are being used in math classrooms. The overall results indicate that the daily use of computers is very low. The highest percentage of daily use for computer use was to solve math problems (16.3%). The next two popular uses for computers in mathematics classrooms on a daily basis were for “to apply learning” and “to practice math drill”. The percentages were 14% and 13% respectively. However, over one-third of the students (34%) reported they never used the computer to practice math drills. The least daily use for computers was to provide “one-on-one instruction”. The reported percentage was 7%. On the other hand, 52% of the students indicated that their teacher never uses the computer for one-on-one instructions. Over 50% of the students reported they never used computers to review math work. Similarly, 52% of the students indicated that their teacher never uses the computer for one-on-one instructions. About 41% of the students indicated that there teacher never uses the computer to show them new topics. Over one-third of the student (35%) reported that they never used the computer to analyze data. Finally, about 30% of the students reported that they never used the computer for graphing math problems.
### Table 3. How computers are used in mathematics classrooms overall frequency

<table>
<thead>
<tr>
<th>Activity</th>
<th>Frequency</th>
<th>How Often</th>
</tr>
</thead>
<tbody>
<tr>
<td>How often do you use computer to review math work</td>
<td>52.4</td>
<td>Never</td>
</tr>
<tr>
<td></td>
<td>22.3</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>8.6</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>7.6</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>9.1</td>
<td>Everyday or almost</td>
</tr>
<tr>
<td>How often do you use computer to solve math problems</td>
<td>28.6</td>
<td>Never</td>
</tr>
<tr>
<td></td>
<td>6.9</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>14.7</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>13.5</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>16.3</td>
<td>Everyday or almost</td>
</tr>
<tr>
<td>How often do you use the computer for graphing in math</td>
<td>30.6</td>
<td>Never</td>
</tr>
<tr>
<td></td>
<td>29.4</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>18.0</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>12.2</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>9.9</td>
<td>Everyday or almost</td>
</tr>
<tr>
<td>How often do you use the computer to practice math drills</td>
<td>33.6</td>
<td>Never</td>
</tr>
<tr>
<td></td>
<td>24.8</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>15.3</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>13.3</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>13.0</td>
<td>Everyday or almost</td>
</tr>
<tr>
<td>How often do you use the computer to analyze data</td>
<td>35.1</td>
<td>Never</td>
</tr>
<tr>
<td></td>
<td>26.4</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>3.3</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>12.1</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>9.2</td>
<td>Everyday or almost</td>
</tr>
<tr>
<td>How often do you use the computer to apply learning</td>
<td>28.0</td>
<td>Never</td>
</tr>
<tr>
<td></td>
<td>26.5</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>18.1</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>13.6</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>13.8</td>
<td>Everyday or almost</td>
</tr>
<tr>
<td>How often does your math teacher use computer for one-on-one instructions</td>
<td>51.6</td>
<td>Never</td>
</tr>
<tr>
<td></td>
<td>22.9</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>8.8</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>6.6</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>9.9</td>
<td>Everyday or almost</td>
</tr>
</tbody>
</table>

*continued on following page*
Table 3. continued

<table>
<thead>
<tr>
<th>How often does your math teacher use the computer to show new topics</th>
<th>41.1</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24.3</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>9.9</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>1.1</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>10.7</td>
<td>Everyday or almost</td>
</tr>
</tbody>
</table>

Table 4. Socioeconomic status and how computers are used in mathematics classrooms percentage of use

<table>
<thead>
<tr>
<th>How often do you use computer to review your math work</th>
<th>38.76%</th>
<th>50.19</th>
<th>56.81</th>
<th>65.61</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27.91</td>
<td>23.74</td>
<td>20.42</td>
<td>16.32</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>11.16</td>
<td>7.34</td>
<td>8.72</td>
<td>6.84</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>9.73</td>
<td>9.27</td>
<td>5.75</td>
<td>5.44</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>12.44</td>
<td>9.46</td>
<td>8.30</td>
<td>5.80</td>
<td>Everyday or almost</td>
</tr>
</tbody>
</table>

Chi. Sq = 96.72, p < .001***

<table>
<thead>
<tr>
<th>How often do you use a computer to solve math problems</th>
<th>21.36</th>
<th>26.19</th>
<th>30.89</th>
<th>36.74</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>29.15</td>
<td>23.81</td>
<td>25.61</td>
<td>28.37</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>15.15</td>
<td>14.97</td>
<td>15.07</td>
<td>13.64</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>12.99</td>
<td>16.67</td>
<td>13.18</td>
<td>11.47</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>21.36</td>
<td>18.37</td>
<td>15.25</td>
<td>9.77</td>
<td>Everyday or almost</td>
</tr>
</tbody>
</table>

Chi. Sq = 70.91, p < .001***

<table>
<thead>
<tr>
<th>How often do you use a computer to graph math problems</th>
<th>30.16</th>
<th>31.90</th>
<th>31.80</th>
<th>28.96</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27.74</td>
<td>28.77</td>
<td>28.75</td>
<td>32.19</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>16.29</td>
<td>16.43</td>
<td>18.30</td>
<td>21.04</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>12.74</td>
<td>13.50</td>
<td>11.93</td>
<td>10.43</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>13.07</td>
<td>9.39</td>
<td>9.15</td>
<td>7.37</td>
<td>Everyday or almost</td>
</tr>
</tbody>
</table>

Chi. Sq = 20.38, p < .060

<table>
<thead>
<tr>
<th>How often do you use a computer to practice math drills</th>
<th>27.41%</th>
<th>32.21</th>
<th>34.61</th>
<th>40.73</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>26.67</td>
<td>21.17</td>
<td>25.05</td>
<td>25.67</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>15.56</td>
<td>16.37</td>
<td>16.25</td>
<td>13.15</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>13.04</td>
<td>17.26</td>
<td>13.00</td>
<td>10.30</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>17.33</td>
<td>12.99</td>
<td>11.09</td>
<td>10.14</td>
<td>Everyday or almost</td>
</tr>
</tbody>
</table>

Chi. Sq = 50.60, p < .001***

<table>
<thead>
<tr>
<th>How often uses computer to analyze data</th>
<th>32.95</th>
<th>33.14</th>
<th>36.46</th>
<th>38.14</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24.75</td>
<td>26.98</td>
<td>24.02</td>
<td>29.56</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>17.54</td>
<td>15.68</td>
<td>19.65</td>
<td>16.61</td>
<td>Less than once a week</td>
</tr>
</tbody>
</table>

continued on following page
never using computers to practice math drills compared to about 27% of the low SES students. About 13% of the second level SES students and 11% of the third level reported using computers everyday or almost everyday to practice math drills. When asked if their math teacher used computers to show them new topics about 13% of the low SES students and 9% of the students in the high SES group indicated this happened everyday or almost everyday. Twelve percent of the third level SES group and 9% of the second level SES students reported their teacher used computers to show them new topics everyday or almost everyday. Equally, about 13% of the low SES students reported using computers everyday or almost everyday to graph math problems. About 7% of the high SES students and 9% of the second and third level SES students reported using computers everyday or almost everyday to graph math problems. 12% of the low SES students and 6% of high SES students reported that they use computers to review their math work everyday or almost everyday. Also, about 66% of the high SES students reported that they never use computers to review their math work compared to 39% of the low SES students. 57% of the third level SES students and 50% of the second level SES

<table>
<thead>
<tr>
<th>How often do you use a computer to apply learning</th>
<th>26.48</th>
<th>27.85</th>
<th>25.53</th>
<th>31.77</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24.41</td>
<td>26.40</td>
<td>28.60</td>
<td>27.10</td>
<td>Rarely</td>
</tr>
<tr>
<td></td>
<td>17.16</td>
<td>17.00</td>
<td>19.77</td>
<td>18.55</td>
<td>Less than once a week</td>
</tr>
<tr>
<td></td>
<td>14.50</td>
<td>13.56</td>
<td>12.86</td>
<td>13.39</td>
<td>Once or twice a week</td>
</tr>
<tr>
<td></td>
<td>17.46</td>
<td>15.19</td>
<td>13.24</td>
<td>9.19</td>
<td>Everyday or almost</td>
</tr>
</tbody>
</table>

Chi. Sq = 27.66, p < .006**
students reported that they never use computers to review their math work.

Twice as many low SES students reported that their teacher used the computer to provide one-on-one instructions everyday or almost everyday compared to students in the high SES group. The reported percentages for the two levels were 11% and 4% respectively. Additionally, 63% of the high SES students compared to 43% of the low SES students indicated that their teacher never used computers for one-on-one instructions. About 50% of the second and third level SES students indicated that their teacher never used computers for one-on-one instructions. Finally, about 11% of the lowest and second level of SES students reported using computers to analyze data everyday or almost everyday compared to only 5% of the students in the high SES group. About 10% of the third level SES students reported using computers to analyze data everyday or almost everyday.

DISCUSSION

The use of technology in the math and science classroom has been a main focus in improving student learning outcomes. Technology not only can provide visual learning in the classroom, it also opens the door to improve higher level thinking skills. The results of the present study indicate that 10th graders use more calculators on a daily basis compared to computers. Moreover, calculator use far outweighs the use of computers in today’s math curriculum. This is also true for the use of the graphing calculator.

The results of the present study suggests that there are important differences in the use of technology in 10th grade mathematics classrooms associated with levels of SES status. Students from low SES families are less likely to use calculators on a daily basis compared to students from high SES families. This also includes the use of the graphing calculator on a daily basis. Low SES students also reported that they were more likely to use computers on a daily basis compared to high SES students. This may dispel previous findings that low SES students have a less opportunity to use computers compared to high SES students. However, the findings do raise a new issue of disparity in calculator use, which there may be need for further exploration.

Another important finding from this study describes students’ overall use of computers. The results from the present study suggest that overall 10th grade students do not use computers often in their mathematics class. This suggests that schools across the country need to do a much better job in integrating technology in the secondary school curriculum. Depending on its use, technology can be very useful in the mathematics classroom. In fact, operating a computer can be simple compared to designing a scientific experiment and solving challenging math problems, that is, as long as the student and the teacher have the tools. For those who do not use the tools, technology can become a diversion or a simple device for entertainment. Access to technological resources in the classroom is only part of the solution (Feldman, 2001). According to Feldman, research leads us to surmise that teachers who feel more prepared to use technology are more likely to use it in instructional activities (Feldman, 2001). If we want to help students from low SES backgrounds, our emphasis should be on providing nourishing support, specifically, providing well-trained teachers and a rigorous curriculum that integrates the use of technology.

Implementing Technology in Urban Schools

One of the major concerns for urban teachers when integrating educational technology in the classroom is the identification of appropriate principles that help achieve high student learning outcomes. Recent research synthesis efforts by the Center for Applied Research in Educational Technology (2006), the International Society for Technology
in Education (2003), and Roblyer (2006) provide principles for appropriate ways to use education technology in urban schools as supported by the education technology research.

Recent research in educational technology (Boster, Meyer, Roberto, & Inge, 2002; Bracewell & Laferriere, 1996; Coley, Cradler, & Engel, 1997; Cradler & Cradler, 1999; Koedinger & Anderson, 1999; Kulik, 2003; White & Frederiksen, 1998) has shown that the effective use of educational technology occurs when the application directly (a) supports the curriculum objectives being assessed; (b) provides opportunities for student collaboration and project/inquiry based learning; (c) adjusts for student ability and prior experience, and provides feedback to the student and teacher about student performance; (d) is integrated throughout the lesson; (e) provides opportunities for students to design and implement projects that extend the curriculum content being assessed; and (f) is used in environments where the organization leadership supports technological innovation.

Some examples of the strategies that have proved successful in influencing student academic performance include students working in pairs on lessons at the computer assisted instruction through social interactions and teamwork (Bracewell & Laferriere, 1996); digital video clips, audio, and graphics to supplement instruction (Boster, Meyer, Roberto & Inge, 2002); mathematics curricula focusing on mathematical analysis of real-world situations supported by computer assisted instructional software program (Koedinger & Anderson, 1999); multimedia creation involving research skills to locate content resources, capability to apply learning to real world situations, organizational skills, and interest in content (Cradler & Cradler, 1999); software programs that allow students to be aware where they are in the inquiry process and to reflect upon their own and other students’ inquiries (White & Frederiksen, 1998); word processing software that utilizes writing prompts (Kulik, 2003); and online feedback among peers who know one another allows learners to feel more comfortable with and adept at critiquing and editing written work (Coley, Cradler, & Engel, 1997).

CONCLUSION

As educators and educational technology professionals demand for research-based evidence about the effectiveness of specific instructional practices has created renewed interest in educational research (Edyburn, Higgins, & Boone, 2005). As a result, there is an urgent need for research that provides evidence about the effectiveness of various educational technology interventions applied to specific subject domains. This is especially needed in mathematics, where we are still trying to get a handle on the achievement gap of students in urban and suburban communities. Research is a critical piece to the puzzle to fully understanding the impact of technology on student achievement. As a result more research is needed on a large scale that will focus and assess some of the following questions not addressed in the present study:

- What instructional design models will help instructors design effective strategies for student learning success specifically in the mathematics classroom?
- What type of professional development experiences do educators need in order to design successful learning experiences for urban students for mathematics?
- How can educators locate innovative and effective interventions for enabling urban students to achieve high academic standards through the use of educational technology in mathematics?
- What is the impact of race, ethnicity, culture, and language as it relates to the use of technology in urban environments?
- What works, for whom, and which conditions does technology impact student achieve-
Re-Examining the Socioeconomic Factors Affecting Technology Use in Mathematics Classroom

ment in urban environments relating to mathematic instructions?

• What new technologies are needed for routine classroom activities for urban learners?
• How can effective strategies be developed to enhance the teaching and student learning process to reflect active learning in an inquiry/project-based environment?

As we look to the future, technology is often viewed as an enticing means of closing the achievement gap. It is seen as a magic bullet to solve all instructional and learning related issues in an educational environment. However, this is not the reality. Statistics on the digital divide have shown are that the use of technology is often based on simple computer-to-student ratios that have little relevance in describing the quality of the technology experience of the use of the intervention in the classroom. Recent advances in educational technology have the potential to significantly enhance the learning and achievement for all students in the urban environment. However, these contributions hold for diverse urban learners suggests unlimited potential for their application in urban schools. Finally, the current accountability environment demands significant attention to questions of efficacy, which must be addressed in the context of using technology to enhance student achievement.

The current literature has implies that innovative approaches used in teaching with technology leaves students with a more effective learning environment that promotes quality teaching and active student learning. Consequently education planners and policy makers must think beyond providing more hardware, software, and connecting schools to the Internet, but instead thinking about keeping urban schools and teachers well-informed and trained in the effective use of technology for educational purposes. One of these investments is meaningless without the other. High-speed connections, complete digital services, and modern computers are basic to every professional workplace and are essential to student learning in the 21st century. However, technology will fail to meet its educational promise if we neglect to equip teachers with the skills they need to understand and use it and transmit this knowledge and skills to the urban learner.

Adhering to these procedures, educators will be able to grow as practitioners in the field and use educational technology to support quality teaching and active student learning. The research examined in this discussion provides teachers with a relevant framework to understand the factors that affect the urban learner and the power of technology in the teaching and learning process can offer the urban learner, thus leading our educational system to fulfill the promise of providing quality teaching and student learning for a more consistent and dynamic educational learning environment that continues to support the ideals and concepts of the great American education system.

REFERENCES


National Telecommunication and Information Administration (1999). *Falling through the net III: Defining the digital divide.* Full report. Re-
Re-Examining the Socioeconomic Factors Affecting Technology Use in Mathematics Classroom


National Telecommunication and Information Administration (2002). *A NATION ONLINE:*


Owens, E.W. & Waxman, H. C. (1994). *The quantity and quality of technology use in urban, suburban, and rural tenth grade mathematics and science classrooms*. In D. Carey, R. Carey, D. A. Willis, & J. Willis (Eds.), *Technology and teacher education annual* (pp. 147-154), Charlottesville, VA: Association for the Advancement of Computing in Education.


Chapter 4.5
Let Them Blog:
Using Weblogs to Advance Literacy in the
K–12 Classroom

David A. Huffaker
Northwestern University, USA

ABSTRACT

This chapter introduces the use of blogs as an educational technology in the K-12 classroom. It argues that blogs can be used to promote verbal, visual, and digital literacy through storytelling and collaboration, and offers several examples of how educators are already using blogs in school. This chapter also reviews issues such as online privacy and context-setting, and ends with recommendations for educators interested in implementing blogs with current curricula.

INTRODUCTION

As Internet technologies continue to bloom, understanding the behaviors of its users remain paramount for educational settings. For teachers, parents, school administrators, and policymakers, learning what types of activities and applications students are using on the Internet is only the surface. Understanding how they are using these applications can provide innovative strategies for learning environments.

Previously, many scholars have explored how Internet users communicate and present themselves online, using computer-mediated communication venues such as e-mail, chat rooms, instant messaging, newsgroups, multi-user domain (MUDs), and personal home pages to examine communication patterns, online identity construction, and even gender differences (Crystal, 2001; Döring, 2002; Greenfield & Subrahmanyam, 2003; Herring, 2000; Lee, 2003; Turkle, 1995; Witmer & Katzman, 1997).

Internet technologies continue to evolve, and it is important for scholars to examine the latest CMC arenas in comparison with past research in hopes of finding new ways to find creative learning solutions and enhance pedagogical method in educational technology. Weblogs, commonly referred to as blogs, represent one of the latest advances in CMC.

A blog can be simply defined as an online journal. Made up of reversed chronological entries...
infused with text, images, or multimedia, blogs embody a place where individual expression and online community development coexist. Not only do the authors, or bloggers, post thoughts and feelings on a Web page for the world to view, but blog readers can comment, creating a dialogue between the blogger and the community he inhabits. Furthermore, bloggers link to other bloggers, creating an interwoven and perhaps interdependent online community of writers and readers. Blog popularity continues to resonate throughout the media, with many scholars suggesting an evolution in individual self-expression, education and research, online journalism, and knowledge management (Alterman, 2003; Blood, 2003; Herring, Scheidt, Bonus, & Wright, 2004; Lasica, 2003; Moore, 2003; Mortenson & Walker, 2002; Oravec, 2002; Pollard, 2003b; Schroeder, 2003).

In a recent survey, Perseus Development Corporation found that among the four million published Weblogs, almost 53% are created by children and adolescents between ages 13-19 (Henning, 2003). With such a strong population of young bloggers, understanding its potential uses within a classroom remains an exciting prospect for educators and parents. Can blogs be used to enhance learning? In what ways can they be used in the classroom?

This chapter hypothesizes that blogs can be effective educational tools in the following ways: (1) they promote verbal and visual literacy through dialogue and storytelling, (2) they allow opportunities for collaborative learning, and (3) they are accessible and equitable to a variety of age groups and developmental stages in education.

In order to evaluate this hypothesis, this chapter will proceed as follows. First, it will provide a thorough explanation of what blogs are, fundamental blog features, how they are used, and the demographics of the blog population. Second, this chapter will define verbal, visual, and digital literacy, and their importance in learning. Third, it will explain how blogs foster literacy through storytelling and peer collaboration. Fourth, this chapter will describe examples where blogs are used in K-12 classrooms, with an emphasis on the previous concepts of storytelling, peer collaboration, and literacy. Finally, this chapter will provide specific recommendations for educators and school administrators interested in implementing blogs in their schools and classrooms.

WHAT IS A BLOG?

Blogs are personal journals written as a reversed chronological chain of text, images, or multimedia, which can be viewed in a Web page and are made publicly accessible on the Web (Huffaker, 2004a; Winer, 2003). As depicted in Figure 1, blogs typically contain text in the form of a “blog post,” offer the ability for readers to comment or provide feedback, contain archives to past blog posts, and link to other blogs and bloggers.1

Blogs are inherently different from personal home Web pages. First, bloggers post entries through manual software, such as a Web browser, or automatic software, which is downloaded off the Internet and used to instantly publish content to the Web. Therefore, bloggers do not need to understand HTML or other Web programming

---

1. Figure 1. An example blog

---

1. Figure 1. An example blog
Let Them Blog

languages to maintain their blogs. Second, the resulting blog page resembles a personal diary with entries sorted by time and date, a much stricter format than personal Web pages of the past.

Blog Features

David Winer, a fellow at Harvard Law School, considered one of the more visible writers and developers of Weblogs, describes some of the more important features of blogs:

- **Weblog Posts**: Weblog posts include a subject title and a body message. Posts can be relatively short in length such as one or two paragraphs, or they could be a long, thoughtful exposé, encompassing a dozen paragraphs. They can comprise a variety of media objects, including text, pictures, graphics, or multimedia, and even popular file formats such as Microsoft Office documents or Adobe PDFs (Winer, 2003). These posts receive a timestamp to denote time and date of the post. Figure 2 demonstrates the format of a typical blog post.

- **Comments**: Readers of a Weblog have an opportunity to respond to a blog post through a comment or feedback link. These comments create a thread, as many readers can comment on a single post (Winer, 2003). They contain a timestamp and are viewable to the public. Bloggers can also respond back to reader comments. Figure 2 highlights the comment section. Blog readers can click “Post a Comment” to add a new comment or “1 Comment” to read previous ones.

- **Archives**: The front page of a blog contains only a certain amount of posts, sometimes two or three and sometimes twenty. For authors who have maintained their blogs for longer periods of time, they can store past blog posts in an accessible, often searchable archive. As depicted in Figure 3, archives can be organized by month, by week, or even by number of posts.

- **Templates**: Another useful feature for Web authors is a set of presentation tools that allow pages to be built from preexisting templates. Blog authors can choose from a variety of graphical layouts, typography, and color schemes. This allows complete customization and a feeling of personalization for bloggers without any sophisticated technical expertise (Winer, 2003). However, some bloggers like to tinker with a Web programming language, to add third-party applications or other bells-and-whistles to their blogs, and the capability to reprogram a blog is available. Figure 4 represents the types of templates Blogger.com, one of the most popular blog-hosted sites, offers.

Figure 2. Blog post with comment link
What is the Blogosphere?

Although some studies suggest the majority of blogs are still highly personalized venues for self-expression (Herring et al., 2004), many blogs contain links to other bloggers, creating an online community that is often referred to as the blogosphere. These blog communities typically share a common purpose or responsibility (Carl, 2003), for example, a group of friends spread out across the world may use blogs as a means to communicate with each other, or a support group might encourage each other’s therapeutic development, or a group of amateur journalists might be spreading news on situations of political strife or violent conflict in an area that a global news agency may be unaware of.

The blogosphere also represents the total population of bloggers, an Internet culture that is continuing to grow. Several reports and Web sites that gather information on blog statistics, which estimate between two and four million blogs, demonstrate the vastness of the blogosphere; many expect this growth to continue (Henning, 2003; Herring et al., 2004; Kumar, Novak, Raghavan, & Tomkins, 2003). Certainly, the acquisition of population blog software or services within online commercial giants such as AOL, Yahoo, and Google suggest that the world may be on the cusp of a blog saturation in Internet culture (Munarriz, 2003; Olsen, 2004; Shirpy, 2003).
Size of the Blogosphere

Blogs continue to be created and abandoned each day, so the exact population figures on the size of the blog population remains in transit. The population size is captured using software that indexes blogs or comes directly from sites that host blogs such as LiveJournal or Blogger. It must also be considered how many blogs are “active” or updated within the last three months, because many blogs become abandoned (Henning, 2003).

- **Blogcount** (http://dijest.com/bc/), a Web site dedicated to understanding how vast the Weblog community is, collects scholarly and industrial reports that discuss technical and demographic issues, from frequency of blog posts to mapping the blog community. By June 2003, Blogcount estimated there were 2.4 to 2.9 million active Weblogs (Greenspan, 2003).

- **BlogCensus** (http://www.blogcensus.net/) uses a software program to search and locate Weblogs and categorize them by language and authoring tool. As of April 2004, the BlogCensus has indexed 1.83 million blogs and estimates at least 1.21 million of these are active.

- **LiveJournal** (http://www.livejournal.com/) publishes daily statistics direct from its servers, citing that there are over 2.9 million “livejournals,” with 1.4 million active, as of April 2004 (specific data can be found at http://www.livejournal.com/stats/stats.txt).

- **Perseus Development Corporation**, which offers enterprise-level surveys regarding software and technology, recently published a white paper that estimates there are 4.12 million blogs, but at least 2.7 million have been temporarily or permanently abandoned (Henning, 2003).

Age and Gender Demographics of the Blogosphere

Most of these surveys suggest that a significant portion of the total blog population is made up of teenagers, and almost evenly partitioned between genders. The largest age distribution of bloggers typically ranges between 13 and 20, which assumes most bloggers are either in early secondary school or beginning college or university. There are some subtle discrepancies in gender within the studies listed below; however, the margins are not far enough apart to suggest severe differences in gender use of blogs.

- Perseus Development Corporation, for instance, finds blogs are dominated by the youth, with 51.5% of all blogs being developed and maintained by ages 13-19. They also find 56% of the total bloggers are female and 44% are male (Henning, 2003).

- A recent academic study of 203 randomly selected Weblogs revealed 54.2% male authors and 45.8% female authors, as well as 40.4% of blog authors being under age 20 (Herring et al., 2004).

- Another academic study of 358 randomly selected blogs found 52% male and 48% female bloggers, and 39% of bloggers are less than 20 years old. However, they also found there are more females than males in the ‘teen’ category (Herring, Kouper, Scheidt, & Wright, 2004).

- BlogCensus randomly sampled 490,000 blogs to find 39.8% male and 36.3% female, with the rest of the blogs unidentifiable in terms of gender (see http://www.blogcensus.net/weblog/).

- Finally, Jupiter Research found that blogging is split evenly among genders (see http://www.jup.com/bin/home.pl) (Greenspan, 2003).
Let Them Blog

Use of Blogs

Understanding the features of the blogs helps distinguish them from other Internet applications, and grasping the size of the blogosphere signifies the popularity of blogs in Internet culture. The next question involves the content of blogs. What are bloggers writing about? The answer not only provides a context for online community interaction, but possible application for educational technology. These can be divided into five areas: (a) personal blogs, (b) community blogs; (c) journalism blogs; (d) education and research blogs; and (e) knowledge blogs.

Personal Blogs

The most popular use of blogs are similar to personal Web sites authored by individuals, which include chronological posts as well as links to other Web sites or Weblogs (Lamshed, Berry, & Armstrong, 2002). Despite the popular notion that Weblogs lean toward external events, or remain highly interlinked with the blogosphere, the majority of Weblogs are still individualistic self-expressions written by one author (Herring et al., 2004).

Community Blogs

Virtual communities develop through the use of a blog (Lamshed et al., 2002). Examples might include a community support group, a site for parents to ask questions and exchange answers, a research community sharing resources and data, or a mirror of an offline community, like a softball team or neighborhood newsletter. Although personal blogs may dominate the blogosphere, the ability for individuals to choose their level of community participation may be another reason for blog popularity, as it allows the blog author to explore individual needs while benefiting from community interactions (Asyikin, 2003). The linkages with other Web sites, people, and ideas even form micro-communities with varying levels of involvement.

Journalism Blogs

The idea of alternative forms of journalism manifesting through Weblogs has received increasing attention in media and scholarship (Alterman, 2003; Blood, 2003; Lasica, 2003). Where is Raed? (http://dear_raed.blogspot.com), for instance, is a blog by an Iraqi that discusses what is happening in Iraq since September 2003. He discloses a candid view of the U.S. occupation, but also introduces readers to fellow Iraqi bloggers. For most, the global news agency is the only link to international affairs—having personal, subjective commentary within a foreign world provides a unique view to outsiders.

A different, but equally unique log is J-Log, which provides community critiques and commentary on current journalism and news. The community not only shares interesting news items, but also poses questions such as, “Is this news fair and balanced?” Perhaps J-Log (http://www.mallasch.com/journalism/) and individual reports such as Raed demonstrate new forms of online journalism; critiques, however, as to the viability of these news sources remain an issue, including the resources and references and even the subjectivity amidst objective journalistic philosophy.

A link to http://blogdex.net/, an MIT Laboratory experiment that captures the fastest-spreading ideas in the blog community, typically results in news headlines as the most contagious information.

Education and Research Blogs

Blogs have been heralded as opportunities to promote literacy in learning by allowing students to publish their own writing, whether it is a journal or story, or even comments on class readings (Kennedy, 2003). For more advanced students,
blogs present the same opportunities: writing and thinking with Weblogs, archiving and analyzing past knowledge, and developing a social network to collaborate and critique (Mortenson & Walker, 2002).

Blogs allow educators and students to interact in the same common space and format (Wrede, 2003). Lamshed et al. (2002) find that students believe blogs are easy to use and navigate, and enthusiastic about learning features such as storing and managing information, communicating, reviewing work before posting, and “keeping on track,” which refers to managing time or monitoring progress (Lamshed et al., 2002). If students are eager to adopt blogs in the classroom, teachers may have an opportunity to also keep on track with students, as communication and interactions are visible and accessible anytime-anywhere.

Several sites explore the use of blogs in education. Weblogg-Ed (http://www.weblogg-ed.com/), maintained by Will Richardson, collects ideas about Weblogging in school settings and facilitates dialogue between teachers. Similarly, Edublog (http://edublog.com/) is an initiative to develop and study blog-based tools for the classroom.

Knowledge Blogs

Similar to education and research, blogs provide opportunities for organizations to manage and share content, as well as communicate across the network. Dave Pollard, Chief Knowledge Officer at Ernst and Young, Inc. and popular writer on the role of blogs in the business, suggests that blogs can be used to store and codify knowledge into a virtual file cabinet. But unlike other content management tools, blogs allow authors to publish in a personal and voluntary way, creating a democratic, peer-to-peer network (Pollard, 2003b). Pollard also suggests companies can increase profitability by designing information architecture to embrace the use of Weblogs (Pollard, 2003a, 2003c).

VERBAL, VISUAL, AND DIGITAL LITERACY

Literacy has always been a focus of learning, especially considering that the foundations of education are grounded in reading and writing. In fact, reading and writing, often referred to as verbal literacy, serves as a benchmark for success in education. Reading and writing are not only important in language arts or humanities; they serve as prerequisites for all academic disciplines, including science and mathematics (Cassell, 2004). Verbal literacy is developed even before children enter school as parents read stories to their children, helping them to understand the relationship between words and pictures, as well as helping to develop narrative structure (Bransford, Brown, & Cocking, 1999; Huffaker, 2004b).

Scholars have recognized how communication also takes place in the form of images and symbols, what is referred to as visual literacy (Gee, 2003). As James Paul Glee (2003) suggests, texts are becoming increasingly multimodal, containing both text and images, and can be recognized in everything from the advertisement to the high school textbook (Gee, 2003). Our society is filled with these multimodal symbols, and their coherence is intrinsic for operating in the modern world. Therefore, reading and writing should not only include words but also images, and the development of both verbal and visual literacy is essential for success inside and outside school walls.

The use of technology represents a third type of literacy, equally ubiquitous and important as the other forms. Digital literacy, sometimes referred to as technological fluency, embodies the idea of using technology comfortably, as one would a natural language (Cavallo, 2000). As users of computers and other digital technology become more fluent, they learn to communicate and express themselves explicitly and eloquently using these tools; in effect, technology becomes innate. Furthermore, just as reading and writing are widely hailed as the building blocks for success
in society, digital literacy becomes necessary for success in the technological world we inhabit.

This is what makes educational technology so exciting—it encourages and advances all three types of literacy: verbal; visual; and digital. Blogs, for example, utilize both textual reading and writing, but also involve the use of graphics in the forms of emoticons, or graphical expressions of emotions, images, and multimedia. Blogs also encourage digital literacy, which is grasped from navigating a graphical user interface, and using computers and the Internet in order to publish content.

Some CMC contexts such as e-mail and instant messaging often utilize short pieces of dialog, informal language, and even altered words from spoken language (Crystal, 2001). Educators might complain that using these forms in a classroom might reduce literacy skills, as language development is often paired with precision and formality. Blogs might combat this issue, as its medium involves longer written passages and can be contextualized by educators to promote formal language skills.

### STORYTELLING AND COLLABORATION

Storytelling is a natural part of adolescent development, and children understand the fundamentals of storytelling in their first three years of life (Bransford et al., 1999). Even as a baby, a parent introduces the child to storytelling via bedtime readings or oral tales. By the time a child reaches age four, he can recall many types of stories, whether fictional or autobiographical (Bransford et al., 1999). The stories develop into more mature narrative as the child gets older, thus storytelling provides a way for children’s language and reading skills to develop (Bransford et al., 1999).

Children advance literacy skills through the practice of telling stories to adults and peers alike (Ryokai, Vaucelle, & Cassell, 2003). With adults, children advance language skills because their partners have even more advanced language skills, and children will adapt (Ryokai et al., 2003; Vygotsky, 1980). With peers of similar age groups, children feel more comfortable to collaborate, and enjoy learning together and building on each other’s knowledge (Ryokai et al., 2003). Because peer relations remain quite important to children (Bullock, 1998), effective group dynamics and learning also results in pro-social behavior through a sense of ownership, compromise, and shared responsibility (Calvert, 1999).

Storytelling does not end in childhood, but continues throughout adolescence and even adulthood. Stories help children and adults alike share experiences and feelings in an engaging and even entertaining way (Denning, 2001). From simple sandbox sketches to dinner party yarns, the importance of stories as a catalyst for conversation and dialogue is clear. Furthermore, storytelling helps create connections between people, to engage people, to captivate them. This captivation occurs on both ends—it is just as fun for the teller of the story as the listener. Storytelling fused with peer collaboration is an excellent way to improve language abilities and advance literacy for children and adolescents (Ryokai et al., 2003).

Blogs have the potential to foster both storytelling and collaboration. First, blogs serve as a natural venue for personalized self-expressions. Like diary entries, blogs take form as stories, anecdotes, or vignettes, similar to the types of oral and written stories people encounter everyday. This includes past and present activities, feelings about oneself or other people, or even hyperbolic or fictitious tales.

Secondly, blogs have technical features that offer the potential to create a dialogue between blog author and blog reader, whether it is a reader response, a question-and-answer sequence, or even general brainstorming. In some cases, blog authors discuss candid feelings and readers respond with encouraging statements, providing therapeutic connections. These comments form
a chain between the author and readers, and in essence, an online community. Communities are also built as bloggers link to each other, creating a group of storytellers that provide individualistic expressions, as well as interactions with each other.

EXAMPLES OF BLOGS IN PRACTICE

Blogs are just beginning to infiltrate classrooms, as educators and school administrators consider blogs as a useful tool for communicating between teachers, schools, students, and parents, and as a way to showcase the work of students (Richardson, 2004). These practices are celebrated on the Internet through communities of educators interested in blogging and education. Will Richardson’s “Weblogg-ed: Using Weblogs and RSS in Education” Web site (http://www.weblogg-ed.com/), for instance, is a useful source of information. His site focuses on best practices, offers a quick start guide for teachers, and links to other bloggers concerned with blogs in education. “Edblogger Praxis” (http://educational.blogs.com/edbloggerpraxis/) is another important Web site which unites educators who blog about their experiences or pedagogical philosophies.

This section will look at examples of blogs in practice, separating them into high school (grades 9-12), elementary and middle schools (grades 4-8), and primary school (grades K-3) in order to contextualize blog use by age and developmental stage, and to provide useful models for educators and school administrators interested in viewing how blogs work at different levels of the school system.

High School Blogs

Will Richardson is a teacher and Supervisor of Instructional Technology and Communications at Hunterdon Central Regional High School in Flemington, New Jersey. He not only initiates school-wide policies for blog use in the classroom, but has found success in using blogs for his own literature and journalism classes. In the literature class, students use blogs to comment and critique on reading assignments. In the journalism class, the students collect news stories to post in their blogs; they also comment on each other’s stories while serving in an editorial role. Therefore, in language arts classes, blogs can be used to generate a discussion using critical analysis, or as a collaborative tool where students comment and edit each other’s work. For Richardson, blogs help students become more aware of their writing, as well as their audience (Kennedy, 2003). Figure 5 represents how blogs can be contextualized to provide assignments for students and allow sections for critique and reflection. Each student has an individual blog, which is linked to a homepage where the teacher assigns work and posts relevant news and other information.

Blogs also help create a community of practice among participating students. They can collaborate with each other and build knowledge. These types of discussions, where ideas are synthesized and news ideas created, may be intrinsic to building critical thinking skills. They may also feel that they are “part of a team,” and that each individual has a responsibility to contribute in order to achieve success for the group. Again, effective group behavior involves shared ownership, compromise, and responsibility, which creates pro-social conduct (Calvert, 1999).

Blogs are accessible to the general public, and Richardson found an important side effect as students began blogging. When the class discussed a new novel, the author of the book accessed the blog and posted an impressive response to the questions students were asking (Richardson, 2004). This may have a profound affect on how students view their own work. For instance, this may give students the impression that their ideas and discussion are important in the real world, providing a sense of empowerment and self-ef-
ficacy. Literature and journalism classes and experts are not the only beneficiaries; political leaders, scientists, artists, and philosophers can also directly participate in an educational blog, making students feel their academic work remains valuable, an important consideration for motivating children to learn.

**Elementary and Middle School Blogs**

Blogs can be utilized in many of the same ways in other grade levels. The Institut St-Joseph in Quebec City, Canada, for instance, uses blogs among fifth and sixth graders in order to practice reading and writing. Implemented by the school principal, Mario Asselin, Institut St-Joseph bloggers use a software program to write about anything and everything that is school related. Similar to Richardson’s realization, the fact that blog posts are being read by anyone in the world has an acute effect on students. They felt empowered as their blogs received comments from total strangers and even Canadian celebrities (Asselin, 2004). Similar to Will Richardson’s work, Figure 6 portrays a homepage blog where projects are assigned and school- or project-related links are provided. Students can comment on homework assignments to the entire class, and still post to their own blog space.

At first, some parents and other readers complained that the language of the student blogs was too informal to be considered good writing, and even contained misspelling and grammar errors. Critics complained that blogs were teaching improper language skills, so Asselin (2004) discussed these challenges with the students. The students came up with a system where each blog post would be reviewed by students and a graphic would be posted alongside the text, which served as a stamp of approval. Suddenly, student writing improved dramatically, as no student wanted mistakes after insisting their quality of writing was excellent (Asselin, 2004).

Similarly, Anne Davis’ *Newsquest* involves fourth through sixth graders who practice writing by commenting on news and events. Students felt that blogging helped them write better stories, extend their vocabulary, and even feel ‘grown up’ knowing their voice is on the Internet. After a period of working with blogs, these students
began to collaborate with Will Richardson’s high school journalists, forming the “Georgia-NJ Connection” blog. Now, older journalists and younger journalists can share ideas, comments, and critiques. Again, this resonates with Vygotsky’s notions that children will adapt to more advanced language skills observed in adults (Ryokai et al., 2003; Vygotsky, 1980). Figure 7 demonstrates how blog information can be scaled down to reach a younger audience.

**Primary School**

Mrs. Dudiak, who teaches second grade at Oakdale Elementary School in Maryland, uses blogs to create writing assignments for her students. She might use a picture such as a waterfall and ask students to write a description, a story, or poetry. She also asks them to discuss favorite books, what ‘types’ of books and stories they like, as well as depictions of books into a “movie in our head,” students reply in the comment section of the blog. This is an excellent example of how blogs can be placed in specific contexts to meet the goals of classroom curriculum. Figure 8 exemplifies how blogs can be simplified to reach even the youngest students. Changing colors, enlarging texts, and contextualizing material to reach the needs of primary students represent easy changes for educators to make.

Lewis Elementary School in Portland, Oregon, uses a blog to showcase K-5 student work, as well as to post important information and news for parents. For instance, photo galleries of artwork are displayed alongside weekly classroom notes from each teacher. This demonstrates how blogs cannot only be effective in practicing and advancing writing and language skills, but also in communicating between teachers, students, and parents. Blogs may give parents more direct exposure to their children’s school life, and thus more opportunities to have an impact on their children’s learning. Little Miami Schools (http://www.littlemiamschools.com) in Morrow, Ohio, also uses a blog to communicate school goals and news, as well as spotlighting the successes of its students and teachers.

In sum, blogs can be used at a variety of age and grade levels to promote reading and writing skills, as well as to provide communication links between students, parents, teachers, and schools. Students can use blogs on their own,
finding stories or news they find interesting, as demonstrated with the classes of Mr. Richardson and Mr. Asselin, or they can have specific assignments designed by the teacher, as seen in the case of Mrs. Dudiak’s class. Furthermore, schools can use blogs to bridge the links between students, teachers, and parents, making demonstration of work and student progress available, as well as school news and schedules. Figure 9 demonstrates how blogs can be used as a central communication hub for teachers, schools, parents, and students to share information.

The next section will address some of the issues and considerations for educators interested in implementing blogs in their schools and classrooms.

**ISSUES AND CONSIDERATIONS**

**School- and District-Wide Considerations**

This chapter has mainly approached blog implementation from the educator perspective. However, it is important to examine the schools and districts that encompass the individual classroom. Not only do school and district administrators remain an intrinsic part to choosing educational technology initiatives, but their choices can have a positive or negative effect on how these choices become applied. For instance, schools and districts have to work within a budget when making choices for technology. Second, they have to choose which hardware and software standards will be equitable across the community. Finally, they have to implement technology within certain ethical, legal, and security issues.

For each of these considerations, blogs still stand as a viable solution. First, most blog software is free and open-sourced, so administrators can host blogs within local technology infrastructure. Second, blogs are delivered via the Web, so standards have already been established—blogs can be implemented anywhere a computer and Internet accessibility is present. Finally, blogs have the option of being available to the public or of being private, an important consideration that will be discussed in the next section.
In sum, the importance of school- and district-wide influence on educational technology implementation cannot be ignored. However, for the same reasons blogs can be a useful tool and practical solution for educators; administrators should also be impressed.

**Issues of Online Privacy**

A key concern for any educator and school system involves student privacy. Cyberstalking or sexual predation is a serious concern for adolescents using the Internet (Earley, 1999; Gilbert, 1995). Many of the blog hosts offer free Web space available to the public on the Internet; information on a blog is publicly accessible. If the content of blogs or identities of students need to remain anonymous or private to the world, blog software and hosts offer other options.

For instance, free blog-hosts such as LiveJournal offer a “Friend’s Only” option, which involves a username and password for entrance. For even more security, blog software such as Moveable-Type, which can be implemented on local school
Web servers, can remain completely in-house, that is, no one outside of school or classroom computers would even have access to the blogs.

While privacy is an important consideration for educators and school administrators, the counter-argument to keeping blogs private would be that making blogs available to the world might generate feedback that will empower bloggers and further develop communication and collaboration. For example, if a classroom is discussing a science project and a professional scientist participates in the discussion, it would add significant value to the learning experience.

Setting Context

One of the primary challenges for using blogs, as with many technologies, in the classroom is the importance of setting a context for learning to occur. While discovery and creativity does abound when adolescents are allowed freedom to explore these CMC venues, some structure needs to be in place to facilitate a learning outcome. For instance, if writing quality is a concern for a teacher, then contextualizing the language to focus on clear and succinct writing skills is a must. Similarly, students have to be encouraged to use the blog on a steady basis and focus on the classroom material.

In sum, contextualizing blogs to the learning experience will serve to produce more exciting and educational blog experiences in the classroom, experiences that parents, administrators, and students alike can observe and reflect upon.

The next section will provide some recommendations for educators interested in implementing blogs in the classroom to promote literacy and learning.

RECOMMENDATIONS

Use free blog software for easy implementation in the classroom. There are several popular Web sites that provide templates, maintenance, and hosting of blogs. Using free blog software makes it an easier decision for educators and school administrators concerned with school budgets and technology requirements. With a computer and an Internet connection, blogs can be set up in minutes. For educators who want to keep blogs inaccessible to the public, most blog hosts offer the opportunity to password-protect blogs. The most popular blog-hosts are:

- **LiveJournal**: [http://www.livejournal.com](http://www.livejournal.com)
- **BlogSpot**: [http://www.blogspot.com](http://www.blogspot.com)
- **Xanga**: [http://www.xanga.com](http://www.xanga.com)
- **MoveableType**: [http://www.moveabletype.org](http://www.moveabletype.org)
- **T-Blog**: [http://www.tblog.com](http://www.tblog.com)

For those who want to implement the blogs on local school servers, there are open source applications such as MoveableType and phpBB which can be altered and used in any expected capacity, although they have technical requirements (documentation are available on these sites).

Open-source applications such as MoveableType and phpBB allow complete customization of the interface and blog application. So educators should feel that blogs can be completely tailored to suit the needs of students and the affordances and constraints of school district policy. The examples of blogs used in this chapter are presented to demonstrate current applications and serve as a baseline for new ideas and implementation. Educators should feel complete flexibility in redesigning blog software for learning experiences.

Fuse emergent learning with curriculum-based assignments. If given the opportunity, students might begin to navigate their own learning paths, a feature that is recommended by the new science of learning as directed by the U.S. National Research Council (Huffaker & Calvert, 2003). Surely, letting students do as they want with blogs will reveal some interesting trends that could be manipulated into important learning
opportunities. However, it may be a better idea to create a hybrid form where students construct their own learning, as well as assignments that teachers feel aid the curriculum. Therefore, students have some flexibility in exploring the medium of blogs, while teachers provide direction and structure in the educational practice.

**Encourage collaboration.** Blogs have great potential as a collaborative environment. As students read and write, they may comment and critique each other’s work. They may even edit each other’s work. Assignments can be created where students have to work as a team to complete the goals, such as a story “train,” or chain of story parts devised by individual students, or literature critiques of different parts of a book, or even a mix of different perceptions on the same experience.

**Develop a system for rating student work before it is published.** One way to maintain the caliber of writing that is expected in formal writing is to create a way to ‘proof’ student work before it goes into the public. Most blogs contain a preview area before publication takes place. The teacher does not necessarily need to be the final editor—students can edit each other’s work and provide the final stamp of approval before the world sees the writing. This may be more difficult at early ages, but the metaphor provides accountability for students, which the Institut St-Joseph found dramatically improved writing quality (Asselin, 2004).

**Create interactions between students and the outside world.** One way to make students feel their work is valuable is to demonstrate how their words are being read outside the classroom. For instance, if parents or even total strangers provide feedback on student work, children and adolescents might become more aware of their writing, as well as their audience (Richardson, 2004). Perhaps even more empowering are experts such as authors, poets, artists, scientists, engineers, or other professionals who may comment on student work or add to the discussion. These linkages are not only useful for learning and achievement, but are not difficult tasks to achieve when using blogs, especially considering their accessibility by the public sphere.

**Link with other educational blogs.** As exemplified in the Georgia-NJ Connection, students using blogs may connect with other classrooms despite geographic constraints, in order to expand the community of practice. Even regardless of age, students can share with other students, providing an online learning community that may become self-motivating and self-sustaining. Blogs provide an easy way to do this with built-in options for linking to other blog communities.

**Get parents involved in reading and participating with blogs.** As previously mentioned, getting parents to provide feedback to student work might be an excellent way for students to feel their work is valuable and even appreciated. Yet this has a two-fold benefit—parents may feel they have more exposure to, and thus more interaction with, their children’s learning. Therefore, both students and parents benefit.

Furthermore, blogs can provide links between schools, classrooms, teachers, students, and parents. Parents who visit a classroom blog may encounter the latest third-grade arts and crafts, see the latest school news, or see what is on for lunch that day. This helps bridge communication between all parties responsible for educating children and adolescents.

Creating linkages between all the classrooms in the school creates an interwoven community where everyone feels connected and accessible. This community of practice may have a paramount effect on increasing the efficiency and success of learning in K-12 educational settings.

**CONCLUSION**

This chapter has sought to demonstrate how blogs can be a useful tool for promoting literacy in the K-12 classroom. Literacy most noticeably takes
form in reading and writing, but visual literacy as well as digital literacy may be just as important for success in education and beyond school walls (Cassell, 2004; Cavallo, 2000; Gee, 2003). As an Internet technology, blogs involve the reading and writing of texts, but also the creation and interpretation of images and multimedia, as well as the need to navigate a graphical user interface, all of which serve to advance skills in verbal, visual, and digital literacy.

Storytelling, a natural part of children’s language development, serves as a catalyst for promoting literacy (Bransford et al., 1999; Huf- faker, 2004b). Children and adults alike use stories to express thoughts, feelings, and experiences in an engaging and entertaining way (Denning, 2001), and storytelling helps create collaborations between the teller and listener. Collaborations, whether between children and adults or among peers, are also important aspects of learning, producing ways to adapt or scaffold learning (Ryokai et al., 2003; Vygotsky, 1980). Blogs, which resemble personal journals, create a perfect environment for sharing stories and other forms of writings, and provide ways for others to participate through feedback or critique on student work.

Currently, blogs in practice highlight that blogs can be used in a broad range of age groups and developmental stages in education. Second graders and high school teenagers alike can use blogs to practice reading and writing, and help develop language skills. Blogs can also be used in a variety of disciplines, from literature and journalism to arts and sciences. Anywhere language is used to promote discussion, blogs serve as an arena where thoughts can be published and made available to everyone for additional feedback. Not only can students use the blogs, teachers, parents, and experts can collaborate alongside them, which may help develop a sense of empowerment and self-efficacy for children and adolescents.

Finally, this chapter has recommended several concepts for educators interested in implementing blogs in the classroom. These concepts embrace collaboration, online community building, interlinking parents with schools, student accountability, and ways to monitor student progress. Because blogs are readily available for free at a selection of blog hosts, implementation into the class does not require a high level of technical expertise or an extravagant school budget. Therefore, the benefits of advancing language skills or bridging communication between teachers, parents, and schools far outweigh the costs.

“Let them blog” is a simple idea—blogs provide a computer-mediated communication context where students can practice and advance different types of literacy, use storytelling to express themselves, and collaborate with peers and adults as they complete assignments. “Letting them blog” is even simpler—blogs are easy to use, free, and extremely accessible for anyone with a computer and Internet access. Teachers may find blogs to be one of the easiest educational technology applications to implement, and since adolescents are already using blogs outside the classroom, it may be an excellent time to connect the two.

ACKNOWLEDGMENTS

The research of this chapter was funded in part by a grant to the Children’s Digital Media Center at Georgetown University (http://cdmc.georgetown.edu) by the National Science Foundation (Grant #0126014). Special thanks to Sandra Calvert for all her mentorship and support.

REFERENCES


ence of Educational Multimedia, Quebec City, Canada.


ENDNOTES

1 Except where noted, all examples of blogs were created by the author in order to avoid any privacy or copyright issues.

2 Blogcensus was developed by the National Institute for Technology and Liberal Education (NITLE), a non-profit organization funded by the Andrew Mellon Foundation.

This work was previously published in Handbook of Research on Literacy in Technology at the K-12 Level, edited by L. Tan and R. Subramaniam, pp. 337-356, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
ABSTRACT

With the development of technology, new roads have been opened in education. An interesting idea is to use computers in teaching and learning procedure. Students will have the opportunity to gain access to information resources in a timeless and limitless way. Teachers will be able to transform their classes in a student-centered environment, avoiding the drawbacks of the traditional teacher-centered model. In this direction, ubiquitous computing has significant advantages. Ubiquitous means that computational devices are distributed into the physical world, giving us boundless access to communication and information channels. Now, knowledge can be built based on collaboration, communication, experimentation, and on students' experiences. Research has shown positive impacts on learning. This chapter deals with issues directly connected to ubiquitous computing, such as its features, types of devices used, and pedagogical goals. The advantages and disadvantages of ubiquitous environments are fully examined and some initiatives are referred.

INSIDE CHAPTER

With this effort we try to cover the subject of ubiquitous or pervasive computing in education. We present important issues related to it. The first is the features of this technology. It is important to see and understand them. We also describe the connection means in a pervasive environment. Devices are basic elements of such systems because all educational activities are based on them. Devices vary from those with small screens to those with larger screens. Another separation may be given based on their computational capabilities. We also deal with the pedagogical goals that must be implemented. It is a crucial part of such efforts because the desired result is to assimilate and efficiently teach the students. We describe the advantages and disadvantages of the referred technology. As we will see, ubiquitous computing offers a lot of interesting advantages, but on the other hand, there are open issues that must be taken into consideration. The last part is devoted to the description of some initiatives that take or took place in universities and schools. We describe some platforms that may be used to
construct learning environments, most of which originated in the USA. Of course, there are common characteristics between these attempts, but their number is large enough to force us to describe only a few. These were randomly selected.

**INTRODUCTION**

New technologies have brought many changes in teaching, and of course in learning. Traditional classrooms are being transformed in order to utilize the advantages of the technology.

Ubiquitous computing (also known as “Pervasive,” “Ambient,” “1 to 1,” or “one to one”) is about distributed computing devices in the environment, with which users are able to gain access to information resources. These devices can be wearable computers, or sensors and computers embedded in everyday objects. On the other hand, ubiquitous computing involves the necessary infrastructures needed to support pervasive computing applications.

Ubiquitous computing integrates technology into the environment, giving the opportunity to users to utilize it anytime and anywhere. It differs from traditional systems where the user is bonded to a computer in a specific place. Now it is possible for a user to utilize the technology without the restriction of place or time.

Ubiquitous computing may provide significant advantages in the application domain of education. It can offer continuous access to a wide range of software, or the Internet, to all students, as well as teachers. As we will see below, the main targets of using pervasive techniques in education are efficiency in teaching and learning, equality between all students as to access to technology, regardless of their economical state, increased student engagement with their lessons, and different approaches according to the students’ needs (Bonifaz & Zucker, 2004).

This chapter is organized as follows. The next section gives information about the examined technology, and the third section describes its basic features. In the fourth section, a full description of the means that are being used in order to help a user connect and utilize ubiquitous facilities is given. The fifth section describes the pedagogical goals of the use of pervasive computing, and in the sixth section, the advantages and disadvantages of the emerged technology are given. In the seventh section, we outline some initiatives in this research area, and in the eighth section, we give a specific case study. Finally, our conclusions are depicted in the last section.

**BACKGROUND**

Ubiquitous computing environments are different from what one traditionally finds in most school settings. It offers to all students and teachers continuous access to a wide range of software, electronic documents, the Internet, and other digital resources for teaching and learning. These initiatives’ goals include increasing economic competitiveness, reducing inequities in access to computers and information between students from wealthy and poor families, and raising student achievement through specific interventions. Other reasons cited for supporting laptop initiatives include improving classroom culture, increasing students’ engagement, making it easier to differentiate instruction according to students’ needs, and solidifying home-school connections (Bonifaz & Zucker, 2004).

The UK government and Scottish executives have listed a number of priorities for ubiquitous education for the 14+ age range. This list is discussed in Smith (2002) and Sutherland (2002). According to authors, the priorities posed are:

- Widening participation of students
- Increasing the diversity of students in education
- Quality and standards through the use of ubiquitous computing
Ubiquitous Computing Applications in Education

• Employability in students’ life
• Accessibility to various resources through the emerged technology
• The globalization of learning and the emergence of alternative HE (Higher Education) suppliers
• Provision of education in more accessible forms on a part-time basis
• Employers wanting more workplace learning
• Professionalisation of teaching
• Student awareness of quality teaching and pastoral care
• Increasing IT literacy of younger students and higher students’ expectations
• Staff handling larger groups

Similar priorities have been posed in other countries recognizing the need to change the traditional educational model that has important disadvantages. To this target, the International Society for Technology in Education (ISTE) published a list (ISTE, 2002) with the most commonly referred conditions to create a learning environment for the use of technology. These factors are:

• Shared vision: This means that the commitment to technology is systemic and continual. Also, there is a proactive leadership and administrative support for the entire system.
• Access: Teachers must have limitless access to current technologies, software and telecommunications networks.
• Skilled educators: The educators that are called to instruct students, who use the technology in their tasks, must be skilled and familiar with the technology and with its use in the teaching procedure. Hence, learning will be more efficient.
• Professional development: Educators must have consistent access to professional development in support of technology use in teaching and learning.
• Technical assistance: Educators and students must have limitless technical support for maintaining and using the technology. Usually a technical helpdesk is established in order to meet these needs.
• Content standards and curriculum resources: Teachers must be knowledgeable in their subject and up to date with the content standards and teaching methods in their courses.
• Student-centered teaching: Ubiquitous access to technology drives the transformation of the traditional teacher-centered system to student-centered. Educators’ methodologies must comply with these new requirements.
• Assessment: There must be continuous assessment of the effectiveness of technology for learning. All the problems and the progress of each program must be recorded and encountered in the programs’ implementation.
• Community support: The community and every school’s partners must provide expertise, support, and resources.
• Support Policies: There must be policies for financing, accessing in networks, and rewards structures that must be in place to support technology in learning.

As we can see, there is an intention to use ubiquitous computing to help students learn in a more productive way. Ubiquitous access to technology by teachers and students has the capacity to transform not only the physical learning environment, but also the learning process, the role of students, and the role of teachers (Fullan, 1999).

Results of some 1 to 1 computing initiatives have shown:

• Improved writing skills.
• Increased student motivation and interest
Ubiquitous Computing Applications in Education

in school.
• Students that are more engaged in their learning.
• Teachers increasingly utilizing project-based and hands-on curriculum and teaching methods.
• Increase of math and science scores in the eighth grade.
• Children born since 1980 process information differently than children born before 1980. They learn best with multisensory input.

A case study is presented later in this chapter. It can be very helpful because it depicts valid results taken from questionnaires and interviews.

There are several research efforts in the field of ubiquitous computing in education. Especially with the spread of the wireless and handheld technology, these efforts become more intensive due to the advantages it can offer to learning process. Current research focuses on the impact of pervasive computing in teachers and students (Bayon et al., 2002). As for the teachers, the goal is to make them use effectively the new technology in their lessons' design in order to produce a student-centered class (Becker & Riel, 2000). This is the basic characteristic of constructivist theory, where students learn based on their collaboration and experimentation. Students build their knowledge based on their experiences and their research.

Regarding students, researchers study the effect ubiquitous learning has on them in terms of when and how they use the technology, how much the technology affects their behaviour, and how the new attempt is received by parents and the school community (Diggs, 2004). Interviews and observations have shown positive effects on students and this way of learning is accepted from the entire school community. There is an enhanced motivation, collaboration, and communication between students. This is also supported in Vahey, Enyedy, and Gifford (2002).

The research results are very optimistic, but there must be careful design and systematic observation to achieve the ultimate goal, which is more productive student learning.

FEATURES

The use of ubiquitous computing in education has characteristics very important in learning. An educational policy can be based on these features in order to achieve a high level of learning. These features are:

• **Information Access**: Students have access to their documents, and also to various information sources from everywhere. All students may search the information needed in order to complete a task. Of course, the initiative is on them. They pose questions and take the results. The final step is to combine the results and extract the final conclusion.

• **Time and Place Immediacy**: This feature is allocated in the place and time in which the information can be reached. Whenever and wherever a student needs to access information is feasible. There are no limits in time or in place. This has the advantage of easy and useful access to resources, increasing the productivity of the work.

• **Interactivity**: Students are able to interact with teachers or experts with a synchronous or asynchronous communication. Hence, they have the opportunity to approach other people's knowledge, without the one teacher's limitation, of the traditional system. They may search, find, and pose questions to specific domain experts, and afterward combine the answers to effectively build their knowledge.

• **Student Activation**: A system in which all students have their own device and work by themselves demands activation. Every student has to work and learn, experiment-
Ubiquitous Computing Applications in Education

ing with software and searching to find the information needed to complete specific tasks.

- **Adaptability:** Learners can get the right information at the right time and at the right place. In this direction, the Intelligent Agent technology may be very useful. Intelligent software may learn from the owner’s habits or instructions and work as its representative, searching and find information. Even further, this software can be used to adapt the information presented to the users based on their learning style.

**CONNECTION MEANS**

When computers first arrived in educational environments, they brought a lot of challenges in the form of space for electrical cords, attachments, peripherals, and other entanglements (McKenzie, 2001). It is obvious that in recent years there was a ‘blast’ in wiring and cabling in school networks.

The ultimate goal of ubiquitous computing in education is the use of a computer for all students and teachers. This means that we need a more flexible way to connect all the devices. The development of the technology in the domain of wireless radio connections, which gives a high speed communication channels, can give us a means to efficiently interconnect devices in large areas.

We believe that a wireless connection will be more appropriate and efficient than a wired network where the users will be bound to specific spots in the area. Using wireless connection, students may have access from anywhere. They may have access in the cafeteria, in the library, in the class or outside.

In McKenzie (2001) authors provide a list of advantages of the wireless technology. They are:

- Ease of movement
- Relaxed fit
- Strategic deployment
- Flexibility
- Cleanliness
- Low profile
- Convenience
- Simplicity
- Speed (especially nowadays)

Synoptically, we can say that the greatest advantage of wireless technology is that it offers continuous access to the networks from anywhere in school. One can see a “wireless” scenario in Figure 1. All the devices can connect to access points and thus can have access to all the network resources. It can be implemented with either infrared rays or wireless Ethernet (e.g., 802.11b, or 802.11g with greater speed). Unfortunately, infrared rays have a disadvantage because they do not penetrate opaque objects. Hence, they may only be used in the classroom. On the other hand, wireless ethernet offers access to the network at a high rate, at any time or place.

The devices that can be used for wireless access vary, from a device with small screen and limited computational capabilities, like Blackberry RIM, to a device with larger screen and increased computational capabilities, like a laptop (Deters, 2001). Some pictures of these devices are shown in Figure 2.

The first kind of device usually allows simple interactions, like message passing, and so forth, but on the other hand, its use is very easy, even for a beginner. It may be used to help beginners enter into this new technology.

Bigger devices like Palm computers, Pocket PCs, or laptops offer increased computational capabilities, giving users the opportunity to utilize more complex software. They have greater memory size and a more powerful processor, and they comprise miniaturization of a desktop machine.
Figure 1. A wireless network example

Figure 2. Portable devices: (1) Laptop, (2) small $100 laptop (3) PocketPC, (4) BlackBerry RIM, (5) handheld laptop, (6) palm computer
The most important matter with respect to devices and their software is compatibility, because students working in groups must exchange messages, files, or even more programs. Their devices must be compatible in order to facilitate the communication between them. Otherwise, there will be a significant problem that may be critical to the entire effort.

PEDAGOGICAL GOALS

Today’s generation of students looks at technology as part of their everyday environment. It is important to understand the difference between today’s life in school in contrast to the past, where students only occasionally used computers. In the future, pupils will own a handheld device, which will be their partner to complete tasks. This means that these devices must be used in a correct manner in order to help them in the learning procedure.

To fully meet the students’ needs, technology should be pervasive—always available. In one-to-one learning, each student has access to a wireless laptop to use at school or at home, enabling communication and collaboration among peers and teachers, and connecting parents to their children’s learning. Educators are provided digital tools to create learning plans, manage educational content, track student progress, and more.

The most important point is that full attention to learning methodologies, with the help of the digital world, must be given. Technology is the mean to learning. A possible mistake will be to pay attention in technology rather than learning. For this reason, policy makers must grapple with several issues concerning appropriate and effective integration of computers into schools (Eagleton, 1999). In the referred list, one can distinguish financial, equality, curricular, and literacy issues.

The ultimate goal of such efforts is to improve the assimilation ability of the students. To this target, there are two critical factors. The first is teachers and their ability to transform their teaching. Teachers must know how to utilize the technology in their courses. They must apply new educational techniques oriented to the use of technology. This procedure needs careful design. It is useless to leave children to play for hours or aimlessly browse the Web. A specific scope must be posed.

The other factor is students. Research indicates that constant exposure to multiple, overlapping sights and sounds has affected the neural pathways of the brains of digital kids (Apple Education—http://www.apple.com/education). In fact, some researchers contend that multisensory input helps kids learn, retain, and use information better. However, the use of technology does not mean that all the students would do their learning with computers. For this reason, it is crucial what type of technology will be distributed to students. Students must be driven to collaborate with their classmates, to use learning software for their courses, to use their devices for communication, or to research with resources useful to complete their tasks. They build their knowledge by experimenting and testing their ideas based on their experiences. The true value of technology for learning lies not in learning to use technology, but in using technology to learn (Educational Research Service, 2001). A technology-using attitude must be cultivated in students in order to enhance their achievement and engagement.

Moreover, computers may help in the direction of the relevancy of schooling (Muir, 2005). Children always wander if what they learn in school may be useful in their future. Laptop initiative makes learning to appear more relevant to life’s requirements. This is because students, with the help of technology, learn based on projects that allow students to use their own learning style. A typical project may be the searching for information in the network and the configuration of it to cover their needs. Hence, students see the
connection between things they learn and their use in real life.

Furthermore, computers are an object that can be used in many ways related to the personal needs of each student. For example, some students may learn better by searching for information, as we described, while others may need multimedia information in order to assimilate the concepts. It is obvious that in such cases the most appropriate and efficient way for learning is working in groups, where all the students can contribute.

Thus, careful design of courses from the teachers’ side may offer significant advantages in learning procedure. Unfortunately, many classrooms do not work that way. Pencil, paper, lecture, textbook, review, and test are still the norm. However, there are important attempts that show us the road to the ubiquitous computing revolution. Some of these tries are presented in Ubiquitous Computing Initiatives.

In summary, ubiquitous computing environments provide portability, flexibility and efficiency for collaborative learning projects (Sottilo, 2003). Teachers have the ability to integrate computers to every aspect of teaching and learning.

**ADVANTAGES – DISADVANTAGES**

In this point, it is necessary to describe all the advantages and disadvantages of the emerged technology. It is critical to identify these issues because it is a key to the road of the embodiment of computers in the learning procedure.

Some advantages are:

- **Teaching efficiency**: With the use of ubiquitous computing, teaching style has to change. Teachers must adapt their methods to the new environment. This active environment gives the opportunity for students to build their knowledge on their own. Therefore, it is imperative that teachers change their lessons plan in order to reflect the new situation. They will have the convenience to adjust the lessons to each student’s needs and the lessons’ goals. Educators can now implement a broader range of new approaches to learning that are proven to be more effective. Also, teachers must encourage and support their students more vigorously, due to the different nature of teaching.

- **Learning efficiency**: One of the most important findings in past years is that student engagement is one of the most critical factors to learning achievement. It is obvious the traditional teaching method has significant drawbacks for its efficiency. A teacher-centered class does not provide the opportunity for students to learn to build their knowledge through collaboration or experiments. Ubiquitous computing offers an environment where students use their devices to access a class from anywhere, collaborate with other students, experiment with appropriate software, and pose questions to their classmates or their teacher. This method is more productive because the knowledge is built on students’ experiences. Students are more active in the course and their engagement is compulsory in order to achieve the goals that are posed to them. Hence, the traditional model is transformed to a student-centered model, where the teacher becomes a consultant to their students. The procedure that we described reflects the method of constructivist pedagogy.

- **Access to technology**: Today, there are many computers in schools placed at labs, but students do not have limitless and timeless access to them. Consider the scenario where all the students have their own device to access teaching resources either from their class or from their homes. In this scenario, all the students are fluent with the use of technology. This is a very attractive idea because in such a class students are technologically equal. This equality can be allocated in the
use of technology. In past years, there was a gap in technology access that translated into inequalities in different fields among people. This is known as “Digital Divide” (Digital Divide.org, 2005). Now, each student has a device and limitless access to digital information augmenting their productivity. The new technology gives more opportunities to students from previously disadvantaged groups. Also, research has shown that giving students a personal device can make learning more efficient (Vahey et al., 2002). Students become more autonomous and self-directed in their learning.

• **Easy to use:** Today’s children are familiar with technology, especially with games. Gameboys™, Playstations™, and even computers are used for gaming. Thereby, it is easy to provide learning devices to students for their use. Of course, instructions must be given to them for the right use of the tools in order to avoid problems from inappropriate use, as we see below. Students want personal devices that, like a toy, are easy to operate.

• **Productivity:** With the emerged technology students can raise their productivity. This is because they build their knowledge based on techniques that are proven by research to give better results. Students become active, taking initiatives and working with their own pace. Also, the interaction with classmates of whom they share common interests plays an important role in the learning process. Finally, we must note that in such cases experimentation is another factor that builds knowledge. Students have boundless access to the Internet or to specific software, and may perform their own tasks with their devices or search information in various resources, such as the Internet or digital databases, and so forth.

• **Portability:** Students, during their time spent in schools, must be familiarized with tools, software, and devices that differ due to the different architecture of each computer. Computers are used mainly in labs devoted to a specific lesson. They spend significant time learning the use of each device and applications that detaches them from the actual learning activity. With their personal devices, they can focus on the learning procedure without worrying about the differences between various tools and applications. They can connect to the network from any place at any time.

Just like all the other domains of technology, ubiquitous computing has some drawbacks. These drawbacks must be taken into consideration by educational policy makers, and by the implementers of this technology, in order to anticipate and avoid problems that may appear in the implementation phase.

• **Cost:** Initiatives such as ubiquitous computing in education, where every student will have their own device in order to perform tasks, have a cost, detected at two points. The first is the financial cost. The provision of a big amount of devices is very costly, but, of course, their prices are reduced as the technology evolves, making them affordable. On the other hand, there is cost in the transformation of the traditional system into the new. A lot of things must change, such as the lesson plan, the teaching style, and the assessments model. This requires much effort from the teachers and the policy makers. It is important to note that the new model requires a careful design of all the teaching parts because in this environment students will be able to act freely.

• **Teachers’ Expertise:** This kind of program needs teachers’ expertise with the technology and with the design of the new kinds of lessons. Teachers must be fluent with this type of education. There is a fear that they will not know how to use the technology and how to design their lessons. Also, there
is a possibility that they may not use the technology in their teaching methodology in an appropriate way. A solution to these problems is to further educate teachers in coping with the new situation.

- **Inappropriate use of Technology:** There is a fear that students may use the electronic devices with which they are provided inappropriately. For this reason measures must be taken. A lot of programs have the mechanisms to control the network access in order to avoid problems. Firewalls are in the front line of security. With them the administrators can check the traffic and discover the cases that are dangerous to students, and even more, to the system.

- **Equipment damage:** Devices are basic elements in ubiquitous computing environments. They are means for access in information sources. Also they are used for communication purposes. Communication and collaboration are the keys to succeed with the pedagogical goals we described in Pedagogical Goals. Hence, problems will arise when technical inefficiencies occur. In such cases, technical support is important for the program’s success; otherwise the participants’ faith may be unsettled.

All these advantages and disadvantages must be taken into consideration from the policy makers, and a specific policy must be set up according to each case’s characteristics.

**UBIQUITOUS COMPUTING INITIATIVES**

This section will present some initiatives and tools originating from all over the world. Of course, it is impossible to present all the efforts. A full list of the schools that apply the laptop initiative is available at [http://www.k12one2one.org/initiatives.cfm?initID=97](http://www.k12one2one.org/initiatives.cfm?initID=97). We present some of them, in order to describe basic characteristics of ubiquitous computing in education. We must note that most of them originate in the United States, where first attempts appeared in the late 90s. Tools for a learning environment’s construction are also presented. All these initiatives are being held in elementary schools or universities. The common characteristic between them is that the majority use portable devices (see *Connection Means*) with wireless access to the network. On their Web sites, evaluations are given for every one of the programs in order to discern their results and identify the advantages and disadvantages. However, it is commonly the conviction that the use of a computing environment in which the students complete their tasks, offers positive remarks in learning procedure, augmenting the enhancement and engagement of the pupils.

**Ulysses**

This is an effort from the University of Laval, in Quebec, Canada (Mantha, 2001). All the students and teachers are provided with a laptop. In the academic year 2001-2002, there were approximately 1800 laptops in the school. Students and teachers invested money to buy the devices. In Ulysses, the laptop is to a large extent a communication tool, and thus there is a need to be able to connect to the network. The connection is established in the classrooms, in the cafeteria, in hallways, and so forth. For financial reasons, the system uses wired connections to the network. There are 17 wired classrooms, of which 11 have the U shape. There are 6 classrooms with movable tables with network connections on the periphery of the room, allowing the grouping of tables for team work. In every classroom there is a console in which one can find a computer and multimedia tools. An intranet was developed to store all the material needed for courses, like exercises, examples, and
Ubiquitous Computing Applications in Education

so forth. Tools used for communication are e-mail, forums, course management utilities, and Athena. Athena is a tool with which the instructor is able to poll the class in a small amount of time and get the students’ opinion about a question. The answers can be short texts, yes/no or a selection in a multiple choice question. Also, the instructor can display the answers on the projection surface at the front of the class. This capability may be very useful in the teaching process.

For user support, two helpdesks were set up. There is a technical helpdesk and a pedagogical. The first is responsible for hardware and software problems. Common issues encountered are computer parts replacement, systems configuration, network connection parameterization, or systems and applications installation. The second helpdesk is responsible for answering questions concerning the use of technology from teachers and students in a pedagogical setting. There are a lot of means to access this support center. One can get support from the desk, from the Web site, over the phone, or via e-mail. This center can help tutors find and apply new means of uses of technology in their lessons.

**eFusion**

This system consists of an initiative of the University of Illinois that originated in the spring of 2002 (Peiper, Warden, Chan, Capitanu, & Kamin, 2005). eFusion is an interactive classroom learning environment where all students and teachers have access to computing devices inside and outside the class. During the lesson, educators present their notes and examples with the help of presentation tools that eFusion provides. Consequently, teachers send their notes to students’ devices through a wireless network and students may make their remarks in presentations and store them for future access. The system supports interactive newsgroups and communication tools that give the opportunity for students to communicate with their instructors or their assistants and take answers during a lecture or asynchronously at a later time.

The most important part of the system is the facility for incorporation of active learning exercises into the lectures, giving instructors a mean to see the comprehension level of the learners in real time. As we understand, it is about a system that has the ability to facilitate teachers to diagnose their students’ needs in a real-time manner during lectures. This is a critical factor because, based on results taken from quizzes or other forms of questions, instructors may change the teaching methodology in order to achieve a better comprehension level.

**WIL/MA Toolkit**

This is a toolkit written in Java and initiated by the University of Mannheim (http://www.lecturelab.de/UCE_4.html). It is a client-server application, where a server provides connection management, user management, and service management. The first supports the establishment of a connection to the users and gives the capability for administrators to monitor the entire communication in and out from the server. The second is used for user identification and authentication. Finally, the service management unit is responsible to inform users of what services are available and also to control data flow.

The system is based on wired connections, but in order to avoid extensive use of cables, students are able to connect to the network with wireless LAN. The devices used by the students are PocketPCs. On the other hand, teachers are able to publish their material, which can be presentations, slides, or files, and furthermore, they can broadcast image and voice to remote places. During their lectures, teachers can also use tools like call-in (spontaneous questions), or quiz and feedback (from the students in real time). Asynchronous tolls are messaging and chat/forum channels.


**University of Texas**

In 2002, the University of Texas initiated a requirement for all teacher education students to obtain a prescribed laptop and software for use throughout their academic preparation (Resta & Tothero, 2005). The goal is the preparation of a new generation of teachers who would be able to use new tools and practices in their teaching. This program’s vision is to prepare the future teachers to enhance their future students to learn in technologically rich classrooms. Its official name is LIFE (Laptop Initiative for Future Educators).

All the students are provided with laptops and software that meet specific requirements, and wireless access to network in their classrooms or throughout the building. The involvement with the technology creates two critical needs. The first is the need to train the students to use the new hardware and software tools. For this reason, in the start of every semester, workshops are offered to familiarize new students with the systems and applications. The second is the need for technical support. For this, a Students Laptop Helpdesk was established. It provides equipment, supplies, and instruction for both students and teachers.

**Maine Public Schools**

This is one of the first programs that try to embody ubiquitous computing in education and is considered the single largest 1 to 1 educational technology program in the world (http://www.maine.gov/portal/education/educ_tech.html). It is a project of the Maine Department of Education. Every 7th and 8th grade student and teacher has a wireless laptop with a rich set of software. Students have the capability to work with word processing, spreadsheet creation, movie processing, e-mail clients, Internet browsing, and other multimedia software. Also, students can take laptops to their home, where they can have access to the network. There are now 37,000 laptops in Maine’s 239 middle level schools. Additionally, there is a small group of educators, educational technologists producing research, documentation, evaluation, and advocacy of 1 to 1 learning. Evaluations have shown that student engagement and attendance are up and behaviour referrals are down (Silvernail, Harris, Lane, Fairman, Gravelle, Smith et al., 2003). The most prominent result is that parents can see this positive effect on their children’s behaviour. The key to the success of this program is to focus on the initiatives as a computer enhanced learning procedures and not merely as technical training. The technology initiatives must focus on teaching and learning, and not on technology, in order to be successful (Schacter, 1995).

Recently, there was an agreement to upgrade the hardware and software tools. The most important change is the wireless network transition to Airport Extreme 801.11g, giving a 54 Mbps speed and improving performance.

**Vermont Public Schools**

This is an initiative developed by the Vermont Department of Education (http://www.k12one2one.org/www.oakgrove.k12.vt.us). Vermont teachers and students have access to ubiquitous technology, which is available on an “as needed” basis. The results of the first year of the program were:

- Teachers were more comfortable with implementing more technology integrated lessons.
- Students were better organized.
- Students created better quality work.
- Students experimented with various tools.
- An impact on student learning because of the full access to the technology.

All the teachers and students are provided with laptops, with which they can have access to the network and to various software tools. They may process text, databases, spreadsheets, video, mu-
Ubiquitous Computing Applications in Education

sic, or even more images. Finally, Web browsing and communication tools are available.

**New Hampshire Schools**

This attempt concerns all the 7th grade students in six of the state’s neediest schools (http://www.nha-dvantage.com/laptop.htm). Approximately 400 students participated in the program the first year. All these students are provided with a laptop and wireless network access. Additionally, cameras, printers, and videoconferencing tools are available to students and teachers. Teachers can control the children’s laptops during classes while students do their assignments. The analysis of the results concerning the first year evaluation indicates many of the most frequently cited benefits of 1 to 1 computing, such as: increased involvement with technology from both teachers and students, increased student engagement, achievement, and motivation, and improved interactions between teachers and students.

**Talbot County Public Schools**

This is a laptop initiative in which all pupils hold a portable device with which they can cope with school tasks. This project has four goals, which are:

- Increased student achievement
- Effective technology for instruction
- Increased student engagement in the learning process
- Improved educational access for and participation by high-risk students and their parents

The program’s drift derives from classroom observations, teacher competencies identified by National Education Technology Standards (NETS) for teachers, lesson plans that utilize technology, and examining students’ work and performances.

Computers were provided to each ninth grade student at the beginning of the 2005-06 school year (http://www.teps.k12.md.us/index.php?page=l_to_1_laptop_initiative). Teachers were given extensive training on the computers and on instructional software.

Students have access to network from any place, in school or at home. Also, a helpdesk is available to students for technical and insurance issues.

**Alexandria City Public Schools**

The Alexandria City Public Schools High School Technology Integration Project provides a 1 to 1 computing environment for all students in grades 9-12 (http://www.acps.k12.va.us/tip/index.php). A laptop is provided to each student, who has the opportunity to keep the device for the duration of the school year. Students have access to a wireless network with which they can access resources and material useful for their tasks. A set of applications is installed on laptops that is useful for common tasks like word processing, as well as a variety of teaching material and applications. Teachers can leave their material in a students’ server, enhancing the electronic exchange of information. In this server, students have personal work folders where they can store their assignments. Printing support is available for students and teachers. Finally, we must note that for the program’s support, there is a help desk that is responsible for repairing damages.

**Hopkins Public Schools**

This pilot program involves approximately 600 students and 30 staff members. The program’s goals are to enhance student’s achievement and engagement. Each student owns a laptop with which to complete his or her assignments. Teachers can give their files and presentations in electronic mode (http://weblogs.hopkins.k12.mn.us/one2one/). Students having wireless access
Ubiquitous Computing Applications in Education

can take the material and, using software like office, Web browsing tools, database, video, music, image processing, and communication software, can cope with their commitments. It is worthwhile to mention that students cannot have access to the Internet from their homes except in the cases that the family has its own Internet provider. Technical support is provided on a full-time basis.

Other Areas in the United States

There are laptop initiatives in 30 states in the United States (http://www.k12one2one.org/initiatives.cfm?initID=97). These states are:

- Alaska
- Arizona
- Arkansas
- California
- Connecticut
- Florida
- Georgia
- Illinois
- Indiana
- Iowa
- Kansas
- Maine
- Massachusetts
- Michigan
- Minnesota
- Missouri
- New Hampshire
- New Jersey
- New Mexico
- New York
- North Carolina
- Ohio
- Oregon
- Pennsylvania
- South Carolina
- Texas
- Vermont
- Virginia
- Washington

These programs either came as district or school initiatives. Like the previously described examples, they are programs that offer wireless access from portable devices to their students and teachers. The goals and the characteristics of these programs are the same as we described in the previous cases.

Kaifu School – Hamburg

The project started in 1999 and lasted until 2003, and it is among the first in Europe (http://www.hamburgmediaschool.com/pages/p118328_v1.html). Since then, the school chose to continue using portable devices in three classes. However, the entire effort is characterized as positive. All the students were provided with wireless laptops and specialized software. Students that continue this program are specialized in organizing themselves and in learning to use technology, which was the goal of the project.

Minervaskolan School

Since 1999, pupils in Minervaskolan have used laptops with wireless access to the network (http://www.minervaskolan.se/). The program’s technical specifications are similar to the others as we described. The most important is that from the project’s evaluation students’ non-attendance approaches to zero and scores to national tests are remarkable. Also, students’ grades appear to be to a continuous increment advocating to the program’s success. Now, Minervaskolan is one of the top schools in Sweden.

CASE STUDY: THE MAINE LEARNING TECHNOLOGY INITIATIVE – TEACHER, STUDENT, AND SCHOOL PERSPECTIVES, MID-YEAR EVALUATION REPORT

In this section, we synoptically show the results taken from the evaluation of the Maine learning technology initiative (Sirvenail et al., 2003). Three core areas are investigated: Teachers and teaching, students and learning, and schools and community. In order to derive useful conclusions for each area, specific questions were posed. The goal of these questions is to identify the impact, the use, and the obstacles of the technology. Therefore, there are three key questions.
Ubiquitous Computing Applications in Education

- How are the laptops being used?
- What are the impacts of the laptops on teachers, students, and schools?
- Are there obstacles to full implementation of the initiative?

We give a short presentation of the evaluation results with regard to each core area.

The evidence in the area of teachers and teaching indicates that the majority of teachers use their laptops in lesson development and classroom instruction. They utilize their devices to prepare and present their lessons. Of course, they agree that the use of the merged technology has positive impacts on their teaching. Now, they are aware of up-to-date information critical to present a meaningful course. On the other hand, they see an obstacle to this effort concerning the technical support. They feel that technical problems and the lack of technical support sometimes limit the use of laptops. This is a critical factor for the program’s success. They also felt that they need time to learn the new technology and its applications and professional development activities designed to help them in the teaching process. Of course, time must be given in order to overcome the limitations.

Students report that they use their laptops to complete tasks, to search for information sources, and to communicate with teachers and classmates. More importantly, they feel an increase in the benefits of their work, both in and out of school. The new technology has increased their interest in school activities, but there are some problems in coping with technical difficulties.

Positive impacts of the application of new technology in learning procedure are recognized from the school community. The difference in children’s interest is visible to their parents. Financial matters are obstacles to this effort, but through creative solutions many schools find ways to minimize the overall cost.

As we have seen, there are many positive impacts in introducing these new technologies to education. Teaching focuses on students’ generating a new environment, where learning comes from collaboration, experimentation, and research. At the same time there are obstacles, especially in technical support, that in time will be eliminated.

CONCLUSION

Ubiquitous computing changes the field of education. It offers a timeless access to information resources and allows learning methods that are difficult to apply in traditional classrooms. Research has shown it has positive impacts on students’ learning. Also, a lot of advantages exist in teaching, and school communities see these positive impacts of the new technology.

In order to integrate computers in education smoothly, careful design is necessary. The focus must be on the pedagogy and not on technology. This is because the final goal is to accomplish a high level of learning. For this reason, we speak about technology enhanced learning, and not about technical training.

Ubiquitous or pervasive computing has significant advantages over traditional teaching methods, and we must work on them to reach the desired results.

REFERENCES


Ubiquitous Computing Applications in Education


APPENDIX I: USEFULURLS

Ubiquitous Computing Information
http://nano.xerox.com/hypertext/weiser/UbiHome.html

Notes on Ubiquitous Computing
http://hci.stanford.edu/cs147/notes/ubicomp.html

Ubiquitous Computing Applications

Ubiquitous Computing Research Activities
http://ubicomp.teco.edu/index2.html

Ubiquitous Computing Evaluation Consortium
http://www.ubiqcomputing.org/lit_review.html

Ubiquitous Computing Education
http://www.ubqed.info/

IEEE Pervasive Computing - A catalyst for advancing research and practice in ubiquitous computing
http://www.computer.org/portal/site/pervasive/
Chapter 4.7
Unraveling the Taste Fabric of Social Networks

Hugo Liu
The Media Laboratory, USA

Pattie Maes
The Media Laboratory, USA

Glorianna Davenport
The Media Laboratory, USA

ABSTRACT

Popular online social networks such as Friendster and MySpace do more than simply reveal the superficial structure of social connectedness—the rich meanings bottled within social network profiles themselves imply deeper patterns of culture and taste. If these latent semantic fabrics of taste could be harvested formally, the resultant resource would afford completely novel ways for representing and reasoning about web users and people in general. This paper narrates the theory and technique of such a feat—the natural language text of 100,000 social network profiles were captured, mapped into a diverse ontology of music, books, films, foods, etc., and machine learning was applied to infer a semantic fabric of taste. Taste fabrics bring us closer to improvisational manipulations of meaning, and afford us at least three semantic functions—the creation of semantically flexible user representations, cross-domain taste-based recommendation, and the computation of taste-similarity between people—whose use cases are demonstrated within the context of three applications—the InterestMap, Ambient Semantics, and IdentityMirror. Finally, we evaluate the quality of the taste fabrics, and distill from this research reusable methodologies and techniques of consequence to the semantic mining and Semantic Web communities.

INTRODUCTION

Recently, an online social network phenomenon has swept over the Web—MySpace, Friendster, Orkut, thefacebook, LinkedIn—and the signs say that social networks are here to stay; they
constitute the social Semantic Web. Few could have imagined it—tens of millions of Web users joining these social network sites, listing openly their online friends and enlisting offline ones too, and more often than not, specifying in great detail and with apparent exhibitionism tidbits about who they are, what music they listen to, what films they fancy. Erstwhile, computer scientists were struggling to extract user profiles by scraping personal homepages, but now, the extraction task is greatly simplified. Not only do self-described personal social network profiles avail greater detail about a user’s interests than a homepage, but on the three most popular sites, these interests are distributed across a greater spectrum of interests such as books, music, films, television shows, foods, sports, passions, profession, etc. Furthermore, the presentation of these user interests is greatly condensed. Whereas interests are sprinkled across hard-to-parse natural language text on personal homepages, the prevailing convention on social network profiles sees interests given as punctuation-delimited keywords and keyphrases (see examples of profiles in Figure 1), sorted by interest genres.

It could be argued that online social networks reflect—with a great degree of insight—the social and cultural order of offline society in general, though we readily concede that not all social segments are fairly represented. Notwithstanding, social network profiles are still a goldmine of information about people and socialization. Much computational research has aimed to understand and model the surface connectedness and social clustering of people within online social network through the application of graph theory to friend-relationships (Wasserman, 1994; Jensen & Neville, 2002; McCallum, Corrada-Emmanuel, & Wang, 2005); ethnographers are finding these networks new resources for studying social behavior in-the-wild. Online social networks have also implemented site features that allow persons to be searched or matched with others on the basis of shared interest keywords.

Liminal semantics. However, the full depth of the semantics contained within social network profiles has been under-explored. This paper narrates one such deep semantic exploration of social network profiles. Under the keyword mediation scheme, a person who likes “rock

Figure 1. Examples of social network profile formats, on Orkut (left) and Friendster (right). Note the similarity of categories between the two.
climbing” will miss the opportunity to be connected with a friend-of-a-friend (foaf) who likes “wakeboarding” because keyword-based search is vulnerable to the semantic gap problem. We envision that persons who like “rock climbing” and “wakeboarding” should be matched on the basis of them both enjoying common ethoi (characteristics) such as “sense of adventure,” “outdoor sports,” and “thrill seeking.” A critic might at this point suggest that this could all be achieved through the semantic mediation of an organizing ontology in which both “rock climbing” and “wakeboarding” are subordinate to the common governor, “outdoor sports.” While we agree that a priori ontologies can mediate, and in fact they play a part in this paper’s research, there are subtler examples where a priori ontologies would always fail. For example, consider that “rock climbing,” “yoga,” the food “sushi,” the music of “Mozart,” and the books of “Ralph Waldo Emerson” all have something in common. But we cannot expect a priori ontologies to anticipate such ephemeral affinities between these items. The common threads that weave these items have the qualities of being liminal (barely perceptible), affective (emotional), and exhibit shared identity, culture, and taste. In short, these items are held together by a liminal semantic force field, and united they constitute a taste ethos.

**What is a taste ethos?** A taste ethos is an ephemeral clustering of interests from the taste fabric. Later in this paper we will formally explain and justify inferring a taste fabric from social network profiles, but for now, it suffices to say that the taste fabric is an \( n \times n \) correlation matrix, for all \( n \) interest items mentioned or implied on a social network (e.g., a book title, a book author, a musician, a type of food, etc.). Taste fabric specifies the pairwise affinity between any two interest items, using a standard machine learning numeric metric known as pointwise mutual information (PMI) (Church & Hanks, 1990). If a taste fabric is an oracle which gives us the affinity between interest items as \( a(x_i, x_j) \), and a taste ethos is some set of interest items \( x_1, x_2, \ldots, x_k \), then we can evaluate quantitatively the strength, or taste-cohesiveness, of this taste ethos. While some sets of interest items will be weakly cohesive, other sets will demonstrate strong cohesion. Using morphological opening and thresholding (Serra, 1982; Haralick, Sternberg, & Zhuang, 1987), standard techniques for object recognition in the image processing field, we can discover increasingly larger sets of strong cohesiveness. The largest and most stable of these we term taste neighborhoods—they signify culturally stable cliques of taste. Visualizing these interconnected neighborhoods of taste, we see that it resembles a topological map of taste space!

Taste neighborhoods and taste ethoi, we suggest, are novel and deep mechanisms for taste-based intrapersonal and interpersonal semantic mediation. Rather than mapping two persons into interest keyword space, or into a priori ontological space, the approach advocated in this paper is to map the two persons first into taste-space, and then to use their shared ethoi and neighborhoods to remark about the taste-similarity of these persons.

**Emergent and implicit semantics.** While our work builds directly upon age-old language modeling techniques in Computational Linguistics, and graph-based associative reasoning in Artificial Intelligence (Collins & Loftus, 1975; Fellbaum, 1998; Liu & Singh, 2004), it is also sympathetic to trends in the Semantic Web literature—away from formal semantics, and toward an embrace of emergent and implicit semantics. In Volume 1 of this journal, Sheth, Ramakrishnan, and Thomas (2005) distinguish between formal, implicit, and powerful (soft) semantics for the Semantic Web movement. Whereas formal semantics must be manually specified, implicit semantics can be readily mined out of the unstructured Web using statistical approaches. Upon further refinement, implicit semantic resources can be transformed into powerful (soft) semantic resources that afford the ability to mediate informal and formal
entities. Related to implicit semantics, emergent semantics is an evolutionary approach to knowledge management (Staab et al., 2002; Aberer et al., 2004) that advocates semantic organization to be shaped from the ground-up, \textit{a posteriori}, and in accordance with the natural tendencies of the unstructured data—such a resource is often called a \textit{folksonomy}. We suggest that online social network profiles give an implicit semantics for cultural taste-space, and that taste fabrics afford a semi-formal, soft semantics appropriate for semantic mediation between informal and formal entities. Finally, arising out of correlation analysis, topological features of the taste fabric—such as taste neighborhoods, identity hubs, and taste cliques—constitute an emergent semantics for taste-space.

**Paper’s organization.** The rest of the paper has the following organization. Section Two lays out a theoretical foundation for representing and computing taste, framed within theories in the psychological and sociological literatures. In particular, it addresses a central premise of our taste-mining approach—“is the collocation of interest keywords within a single user’s profile meaningful; how does that tell us anything about the fabric of taste?” The section titled “Weaving the Taste Fabric” narrates the computational architecture of the implementation of taste fabric, including techniques for ontology-driven natural language normalization, and taste neighborhood discovery. The section “What Is a Taste Fabric Good For?” describes three semantic functions of a taste fabric—semantically flexible user modeling, taste-based recommendation, and interpersonal taste-similarity—within the context of three applications—InterestMap (Liu & Maes, 2005a), Ambient Semantics (Maes et al., 2005), and IdentityMirror. The following section evaluates the quality of the taste fabric by examining its efficacy in a recommendation task, and also entertains an advanced discussion apropos related work and reusable methodologies distilled from this research. The final section in the paper is the conclusion.

**THEORETICAL BACKGROUND**

This section lays a theoretical foundation for how taste, identity, and social network politics are approached in this work. For the purposes of the ensuing theoretical discussion, social network profiles of concern to this project can be conceptualized as a bag of interest items which a user has written herself in natural language. In essence, it is a self-descriptive free-text user representation, or harkening to Julie Andrews in \textit{The Sound of Music}, “these are a few of my favorite things.” A central theoretical premise of mining taste fabric from social network profiles by discovering latent semantic correlations between interest items is that “the collocation of a user’s bag of interest items is meaningful, structured by his identity, closed within his aesthetics, and informs the total space of taste.” Next, the paper argues that a user’s bag of interests gives a true representation of his identity, and enjoys unified ethos, or, \textit{aesthetic closure}. This is followed with a section which plays devil’s advocate and betrays some limitations to our theoretical posture. The section theorizes a segregation of user’s profile keywords into two species—identity-level items vs. interest-level items. This distinction has implications for the topological structure of the taste fabric.

**Authentic Identity and Aesthetic Closure**

In the wake of this consumer-driven contemporary world, the proverb “you are what you eat” is as true as it has ever been—we are what we consume. Whereas there was a time in the past when people could be ontologized according to social class, psychological types, and generations—the so-called demographic categories—today’s world
Unraveling the Taste Fabric of Social Networks

is filled with multiplicity, heterogeneity, and diversity. The idea that we now have a much more fine-grained vocabulary for expressing the self is what ethnographer Grant McCracken, echoing Plato, calls \textit{plenitude} (McCracken, 1997). In a culture of plenitude, a person’s identity can only be described as the sum total of what she likes and consumes. Romantic proto-sociologist Georg Simmel (1908/1971) characterized identity using the metaphor of our life’s materials as a broken glass—in each shard, which could be our profession, our social status, our church membership, or the things we like, we see a partial reflection of our identity. These shards never fully capture our individuality, but taken together, they do begin to approach it. Simmel’s fundamental explanation of identity is Romantic in its genre. He believed that the individual, while born into the world as an unidentified \textit{content}, becomes over time reified into identified \textit{forms}. Over the long run, if the individual has the opportunity to live a sufficiently diverse set of experiences (to ensure that he does not get spuriously trapped within some local maxima), the set of forms that he occupies—those shards of glass—will converge upon an authentic description of his underlying individuality. Simmel believes that the set of shards which we collect over a lifetime sum together to describe our true self because he believes in authenticity, as did Plato long before him, and Martin Heidegger after him, among others.

While Simmel postulated that earnest self-actualization would cause the collection of a person’s shards to converge upon his true individuality, the post-Freudian psychoanalyst Jacques Lacan went so far as to deny that there could be any such true individual—he carried forth the idea that the ego (self) is always constructed in the Other (culture and world’s materials). From Lacan’s work, a mediated construction theory of identity was born—the idea that who we are is wholly fabricated out of cultural materials such as language, music, books, film plots, etc. Other popular renditions of the idea that language (e.g., ontologies of music, books, etc.) controls thought include the Sapir-Whorf hypothesis, and George Orwell’s \textit{newspeak} idea in his novel \textit{1984}. Today, mediated construction theory is carried forth primarily by the literature of feminist epistemology, but it is more or less an accepted idea.

At the end of the day, Simmel and Lacan have more in common than differences. Csikszentmihalyi and Rochberg-Halton (1981), succeed in the following reconciliation. Their theory is that the objects that people keep in their homes, plus the things that they like and consume, constitute a “symbolic environment” which both echoes (Simmel) and reinforces (Lacan) the owner’s identity. In our work, we take a person’s social network profile to be this symbolic environment which gives a true representation of self.

If we accept that a user profile can give a true representation of self, there remains still the question of closure. Besides all being liked by a person, do the interests in his bag of interests have coherence amongst themselves? If it is the case that people tend toward a tightly unified ethos, or \textit{aesthetic closure}, then all the interests in a person’s bag will be interconnected, interlocked, and share a common aesthetic rationale. If there is aesthetic closure, then it will be fair for our approach to regard every pair of interest co-occurrences on a profile to be significant. If we know there is not any closure, and that people are more or less arbitrary in what interests they choose, then our approach would be invalid.

Our common sense tells us that people are not completely arbitrary in what they like or consume, they hold at least partially coherent systems of opinions, personalities, ethics, and tastes, so there should be a pattern behind a person’s consumerism. The precise degree of closure, however, is proportional to at least a person’s ethicalness and perhaps his conscientiousness. In his \textit{Ethics} (350 B.C.E.), Aristotle implied that a person’s possession of ethicalness supports closure because ethics lends a person \textit{enkrasia} or continence, and thus the ability to be consistent. Conscientiousness, a
Unraveling the Taste Fabric of Social Networks

dimension of the Big Five personality theory (John, 1990), and perhaps combined with neuroticism, a second dimension in the same theory, would lead a person to seek out consistency of judgment across his interests. They need not all fall under one genre, but they should all be of a comparable quality and enjoy a similarly high echelon of taste. Grant McCracken (1991) coined the term Diderot Effect to describe consumers’ general compulsions for consistency—for example, John buys a new lamp that he really loves more than anything else, but when he places it in his home, he finds that his other possessions are not nearly as dear to him, so he grows unhappy with them and constantly seeks to upgrade all his possessions such that he will no longer cherish one much more than the others. Harkening to the Romantic hermeneutics of Friedrich Schleiermacher (1809/1998), we might seek to explain this compulsion for uniformity as a tendency to express a unified emotion and intention across all aspects of personhood. Indeed, McCracken termed this uniformity of liking the various things we consume, Diderot Unity. Diderot Unity Theory adds further support to our premise that for the most part, a person’s bag of interests will have aesthetic closure.

Upper bounds on Theoretical Ideal

From the above discussion, we could conclude a theoretically ideal situation for our taste-mining approach – (1) a user’s bag of interests is an authentic and candid representation of what the user really likes, and (2) none of the interests are out-of-place and there is strong aesthetic closure and share taste which binds together all of the interests in the bag. Here, we raise three practical problems which would degrade the theoretically ideal conditions, thus, constituting an upper bound; however, we suggest that these would degrade but not destroy our theoretical premise, resulting in noise to be introduced into the inference of the taste fabric.

A first corruptive factor is performance. Erving Goffman (1959) posed socialization as a theatrical performance. A social network is a social setting much like Goffman’s favorite example of a cocktail party, and in this social setting, the true self is hidden behind a number of personae or masks, where the selection of the mask to wear is constrained by the other types of people present in that setting. Goffman says that we pick our mask with the knowledge of those surrounding us, and we give a rousing performance through this mask. In other words, the socialness of the social network setting would rouse us to commit to just one of our personae, and to give a dramatic performance in line with that persona. Performance might strength aesthetic closure, but it could also be so overly reductive that the bag of interests no longer represent all of the aspects of the person’s true identity.

A second corruptive factor is publicity. In her ethnographic review of the Friendster social networking site, Danah Boyd (2004) raises concerns over the quality and truth of profiles in light of the fact that a personal profile is public, not only to strangers, but also to one’s high school friends, college friends, professors, ex-girlfriends, and coworkers alike. Because social networking sites generally make a profile visible to all these different social circles at once, Boyd suggests that some users are cowed to the fear of potentially embarrassing exposure—for example, teacher exposing to his students, or teenager exposing to his mother. As a result, users may be cowed into a lowest-common-denominator behavior, sanitizing the personal profile of all potentially embarrassing, incriminating, or offensive content.

Finally, a third corruptive factor also raised by Boyd, concerns the integrity and timeliness of social networks themselves. Boyd claims that Friendster profiles and friend connections are not frequently updated, leading to stale information which could distort the taste fabric if we were interested in looking at the temporal animation of the fabric. Boyd also writes about a phenom-
enon known as Fakesters—the creation of bogus user profiles such as for celebrities. However, the scope of Fakesters is arguably limited, and since Fakesters are chiefly imitations of actual people, aesthetic closure should still be observed and learning over Fakester profile examples should not greatly compromise the integrity of the meaning implied by the taste fabric.

**Identity Keywords vs. Interest Keywords**

While each social network has an idiosyncratic representation, the common denominator across all the major web-based social networks we have examined is the representation of a person’s broad interests (e.g., hobbies, sports, music, books, television shows, movies, and cuisines) as a set of keywords and phrases. But in addition, more than just interests, higher-level features about a person such as cultural identities (e.g., “raver,” “extreme sports,” “goth,” “dog lover,” “fashionista”) are also articulated via a category of special interests variously named, “interests,” “hobbies & interests,” or “passions.”

As shown in the web page layout of the personal profile display (Figure 1), the special interests category appears above the more specific interest categories. We suggest that this placement encourages a different conceptualization for the special interests category—as a container for descriptions more central to one’s own self-concept and self-identification. Of course, syntactic and semantic requirements are not enforced regarding what can and cannot be said within any of these profile entry interfaces, but based on our experiences, with the exception of those who are intentionally tongue-and-cheek, the special interests category is usually populated with descriptors more central to the self than other categories. For example, a person may list “Nietzsche” and “Neruda” under the “books” category, and “reading,” “books,” or “literature” under the special interests category. In the profile normalization process, identity descriptors are inferred from descriptors listed under the special interests category (e.g., “dogs” → “Dog Lover,” “reading” → “Book Lover,” “deconstruction” → “Intellectual”).

Theoretically speaking, it is desirable to have two different granularities of description for a person. Identity descriptors are more general and constitute a far smaller ontology than interest descriptors, thus, the resulting effect is to create a taste fabric structured according to a hub-and-spoke topology. Identity descriptors serve as hubs and interest descriptors serve as spokes. The advantages to such an organization are revealed in a later section on applications, and in the evaluation of the taste fabric in a recommendation task.

Having established a theoretical premise for mining taste fabric from social network profiles, and having argued for identity descriptors to be separate from interest descriptors, the following section dives into the architecture and techniques of the taste fabric implementation.

**WEAVING THE TASTE FABRIC**

The implementation of the taste fabric making system was completed in approximately 3,000 lines of Python code. As depicted in Figure 2, the architecture for mining and weaving the taste fabric from social network profiles can be broken down into five steps: (1) acquiring the profiles from social networking sites, (2) segmentation of the natural language profiles to produce a bag of descriptors, (3) mapping of natural language fragment descriptors into formal ontology, (4) learning the correlation matrix, and (5) discovering taste neighborhoods via morphological opening, and labeling the network topology. The following subsections examine each of these phases of processing more closely. A condensed description of the mining process, sans neighborhood discovery, can be found in Liu and Maes (2005a).
Acquiring Profiles from Social Networking Sites

The present implementation of the taste fabric sources from a one-time crawl of two Web-based social network sites, which took place over the course of six months in 2004. The shortcoming of this approach is that we were only able to mine 100,000 personal profiles from the sites, and only approximately 80% of these profiles contained substantive content because about 20% of users elected to not make their profile details publicly visible to our robotic crawler. Also, the one-time crawls prevents us from being able to engage in more interesting dynamic tracking of profiles which would potentially allow us to animate taste fabrics through time. At press time, we are in discussions with two social network Web sites to gain research access to their user profiles, which should allow for the implementation that we discuss in this paper to be deployed on a larger scale.

At every step of mining, we are careful not to compromise the anonymity or personal information of social network users. In fact, in the end product, all traces of individual users, as well as their idiosyncratic speech, are purged from the taste fabric. From our 100,000 seed profiles, we used only the text of the categorical descriptors and none of the personal information including names and screen names. We chose two social networks rather than one, to attempt to compensate for the demographic and usability biases of each. One social network has its membership primarily in the United States, while the other has a fairly international membership. Both however, have nearly identical descriptor categories, and both sites elicit users to specify punctuation-delimited descriptors rather than sentence-based descriptors. One cost to mining multiple social networks is that there is bound to be some overlap in their memberships (by our estimates, this is about 15%), so these twice-profiled members may have disproportionately greater influence on the produced fabric.
Segmenting Profiles

Once profile texts are acquired, these texts need to be segmented. First, texts are easily segmented based on their interest categories. Recall in Figure 1 that texts are distributed across templated categories, e.g., passions/general interests, books, music, television shows, movies, sports, foods, “about me.” Experience with the target social network websites tell us that most users type free-form natural language text into “about me,” and natural language fragments for the specific interest categories. For the passions/general interest category, text is likely to be less structured than for specific interest categories, but still more structured than “about me.” Perhaps this is due to the following psychology—for specific interests, it is clear what the instances would be, e.g., film names, director names, and film genres for the films category, yet for the general interest category, the instance types are more ambiguous—so that field tends to elicit more idiosyncratic speech.

For each profile and category, its particular style of delimitation is heuristically recognized, and then applied. Common delimitation strategies were: comma-separated, semicolon-separated, stylized character sequence-separated (e.g. “item 1 \.. item 2 \..”), new line—separated, commas with trailing ‘and’, and so on. Considering a successful delimitation as a category broken down into three or more segments, approximately 90% of specific categories were successfully delimited, versus about 75% of general categories. We did not attempt to segment “about me.” Unsegmentable categories were discarded.

Ontology-Driven Natural Language Normalization

After segmentation, descriptors are normalized by mapping them into a formal ontology of identity and interest descriptors (Figure 3). Newly segmented profiles are represented as lists containing casually-stated natural language fragments referring to a variety of things. They refer variously to authorships like a book author, a musical artist, or a filmmaker; to genres like “romance novels,” “hip-hop,” “comedies,” “French cuisine”; to titles like a book’s name, an album or song, a television show, the name of a sport, a type of food; or to any combination thereof, e.g., “Lynch’s Twin Peaks,” or “Romance novels of Danielle Steele.” To further complicate matters, sometimes only part of an author’s name or a title is given, e.g., “Bach,” “James,” “Miles,” “LOTR,” “The Matrix trilogy.” Then of course, the items appearing under the general interest categories can be quite literally anything.

Figure 3 presents the ontology of descriptor instance types for the present taste fabric. At the
top-level of the ontology are six specific interest categories plus one general interest category (i.e., “identities”). Also, as shown, there are roughly 25 second-level ontological types. There are a total of 21,000 recognizable interest descriptors, and 1,000 recognizable identity descriptors, sourcing from ontologies either scraped or XML-inputted from The Open Directory Project (dmoz), the Internet Movie Database (imdb), TV Tome, TV Guide, Wikipedia, All Music Guide, AllRecipes, and The Cook’s Thesaurus. Figure 3 only lists the primary sources, and lists them in order of descending saliency. The diversity and specificity of types ensures the maximal recognition capability over the free-form natural language in the profiles.

Ontologizing identity. The ontology of 1,000 identity descriptors required the most intensive effort to assemble together, as we wanted them to reflect the types of general interests talked about in our corpus of profiles; this ontology was hand-engineered out of a few nomenclatures and folksonomies—most prominently Wikipedia’s extensive list of subcultures and The Open Directory Project’s hierarchy of subcultures and hobbies. We also generated identity descriptors in the form “(blank) lovers” where blank was replaced with major genres in the rest of our ontology, e.g., “book lovers,” “country music lovers,” etc. Some profiles simply repeat a select subset of interest descriptors in the identity descriptors category, so having the “(blank) lovers” template would facilitate the system in recognizing these examples. The mapping from the general interest category into the identity descriptors ontology is far more indirect a task than recognizing specific interests because the general interest category does not insinuate a particular ontology in its phrasing. Thus, to facilitate indirect mapping, each identity descriptor is annotated with a bag of keywords which were also mined out from Wikipedia and The Open Directory Project—so for example, the “Book Lover” identity descriptor is associated with, inter alia, “books,” “reading,” “novels,” and “literature.” Because we employed two parallel mechanisms for identity descriptors, i.e., cultures versus “(blank) lovers,” we cannot be completely assured that these do not overlap—in fact, they are known to overlap in a few cases, such as “Book Lovers” and “Intellectuals” or “Indie Rock Music Lovers” (genre of music) and “Indie” (subculture). Most cases of overlap, however, are much more justified because the cultural lexicon, just as natural language, cannot be flattened to a canon. Perhaps the most debatable choice we made was—for the sake of bolstering recognition rates—up-casting descriptors until they could be recognized in the identity ontology. For example, while “Rolling Stones” is not in the ontology of identity descriptors, we automatically generalize it until it is recognized, or all generalizations are exhausted—in the case of “Rolling Stones,” it is up-cast into “Classic Rock Music Lovers.”

Popularity-driven disambiguation. To assist in the normalization of interest descriptors, we gathered aliases for each interest descriptor and statistics on the popularity of certain items (most readily available in The Open Directory Project) that the system uses for disambiguation. For example, if the natural language fragment says simply “Bach,” the system can prefer the more popular interpretation of “JS Bach” over “CPE Bach.”

Situated semantic recognition. Once a profile has been normalized into the vocabulary of descriptors, they are relaxed semantically using a spreading activation (Collins & Loftus, 1975) strategy over the formal ontology, because more than simply being flat wordlists, the ontological instances are cross-annotated with each other to constitute a fabric of metadata. For example, a musical genre is associated with its list of artists, which in turn is associated with lists of albums, then of songs. A book implies its author, and a band implies its musical genre. Descriptors generated through metadata-association are included in the profile, but at a spreading discount of 0.5 (read: they only count half as much). This ensures that
when an instance is recognized from free-form natural language, the recognition is situated in a larger semantic context, thus increasing the chances that the correlation learning algorithm will discover latent semantic connections.

In addition to popularity-driven disambiguation of, e.g., “Bach” into “JS Bach,” we also leverage several other disambiguation strategies. Levenshtein (1965/1966) edit distance is used to handle close misspellings such as letter deletions, consecutive key inversions, and qwerty keyboard near-miss dislocations, e.g., “Bahe” into “Bach.” Semantically empty words such as articles are allowed to be inserted or deleted for fuzzy matching, e.g. “Cardigans” into “The Cardigans” (band).

Using this crafted ontology of 21,000 interest descriptors and 1,000 identity descriptors, the heuristic normalization process successfully recognized 68% of all tokens across the 100,000 personal profiles, committing 8% false positives across a random checked sample of 1,000 mappings. Here, “tokens” refer to the natural language fragments outputted by the segmentation process; a recognition is judged successful if after stripping away semantically empty words, the token finds correspondence with an instance in the ontology, while remaining within the heuristically-specified tolerances for misspelling and popularity-driven disambiguation. We suggest that this is a good result considering the difficulties of working with free text input, and with an enormous space of potential interests and identities.

**Correlation: Weaving the Raw Fabric**

From the normalized profiles now each constituted by normalized identity and interest descriptors, correlation analysis using classic machine learning techniques reveals the latent semantic fabric of interests, which, operationally, means that the system should learn the overall numeric strength of the semantic relatedness of every pair of descriptors, across all profiles. In the recommender systems literature, our choice to focus on the similarities between descriptors rather than between user profiles reflects an item-based recommendation approach such as that taken by Sarwar et al. (2001).

Technique-wise, the idea of analyzing a corpus of profiles to discover a stable network topology for the interrelatedness of interests is similar to how latent semantic analysis (Landauer, Foltz, & Laham, 1998) is used to discover the interrelationships between words in the document classification problem. For our task domain though, we chose to apply an information-theoretic machine learning technique called pointwise mutual information (Church & Hanks, 1990), or PMI, over the corpus of normalized profiles. For any two descriptors \( f_1 \) and \( f_2 \), their PMI is given in equation (1). The probability of a descriptor, \( \Pr(f) \), is defined here as the frequency of global occurrences of \( f \) divided by the summed frequency of global occurrences for all descriptors.

\[
PMI(f_1, f_2) = \log_2 \left( \frac{\Pr(f_1, f_2)}{\Pr(f_1) \Pr(f_2)} \right)
\]

Looking at each normalized profile, the learning program judges each possible pair of descriptors in the profile as having a correlation, and updates that pair’s PMI. What results is a 22,000×22,000 matrix of PMIs, because there are 21,000 interest descriptors and 1,000 identity descriptors in the ontology. After filtering out descriptors which have a completely zeroed column of PMIs, because there are 21,000 interest descriptors and 1,000 identity descriptors in the ontology. After filtering out descriptors which have a completely zeroed column of PMIs, and applying thresholds for minimum connection strength, we arrive at a 12,000×12,000 matrix (of the 12,000 descriptors, 600 are identity descriptors), and this is the raw interest fabric. This is too dense to be visualized as a semantic network, but we have built less dense semantic networks by applying higher thresholds for minimum connection strength, and this is the reason why small clusters seem to appear in the InterestMap taste fabric visualization.
Criticisms and limitations. A common critique heard about our approach is one that questions the efficacy of using the PMI metric for association. It has been suggested that we should look at collocations of greater rank than binary. Following our initial InterestMap publication, we extended the work by using morphological opening plus thresholding, as is done in image processing, to try to discover larger blocks of collocations which we call neighborhoods. This is to be discussed imminently. Additionally, another suggestion we are considering is negative collocation, that is, the collocation of a descriptor’s absence with other descriptors. This would address an apparent flaw of pointwise mutual information, which is that it “overvalues frequent forms” (Deane, 2005), and would shed a new interpretation on the Semiotician Ferdinand de Saussure’s structuralist enunciation that meaning must be ‘negatively defined’ (1915/1959).

Looking at Topological Features

The raw fabric has two extant topological features worthy of characterization—identity hubs and taste cliques. In addition, we describe what we believe to be a novel application of mathematical morphology (Serra, 1982; Haralick, Sternberg, & Zhuang, 1987) in conjunction with spreading activation (Collins & Loftus, 1975) to discover the taste neighborhoods we hinted at in Section 1.

Identity hubs behave like seams in the fabric. Far from being uniform, the raw fabric is lumpy. One reason is that identity hubs “pinch” the network. Identity hubs are identity descriptor nodes which behave as “hubs” in the network, being more strongly related to more nodes than the typical interest descriptor node. They exist because the ontology of identity descriptors is smaller and less sparse than the ontology of interest descriptors; each identity descriptor occurs in the corpus on the average of 18 times more frequently than the typical interest descriptor. Because of this ratio, identity hubs serve an indexical function. They give organization to the forest of interests, allow interests to cluster around identities. The existence of identity hubs allows us to generalize the granular location of what we are in the fabric, to where in general we are and what identity hubs we are closest to. For example, it can be asked, what kinds of interests do “Dog Lovers” have? This type of information is represented explicitly by identity hubs.

Taste cliques as agents of cohesion. More than lumpy, the raw fabric is denser in some places than in others. This is due to the presence of taste cliques. Visible in Figure 5, for example, we can see that “Sonny Rollins,” is straddling two cliques with strong internal cohesion. While the identity descriptors are easy to articulate and can be expected to be given in the special interests category of the profile, tastes are often a fuzzy matter of aesthetics and may be harder to articulate using words. For example, a person in a Western European taste-echelon may fancy the band “Stereolab” and the philosopher “Jacques Derrida,” yet there may be no convenient keyword articulation to express this. However, when the taste fabric is woven, cliques of interests seemingly governed by nothing other than taste clearly emerge on the network. One clique for example, seems to demonstrate a Latin aesthetic: “Manu Chao,” “Jorge Luis Borges,” “Tapas,” “Soccer,” “Bebel Gilberto,” “Samba Music.” Because the cohesion of a clique is strong, taste cliques tend to behave much like a singular identity hub, in its impact on network flow. In the following Section, we discuss how InterestMap may be used for recommendations, and examine the impact that identity hubs and taste cliques have on the recommendation process.

Carving Out Taste Neighborhoods with Mathematical Morphology

From the raw fabric, another step of processing is needed to reveal taste neighborhoods—patches of taste cohesion that are larger than taste cliques.
Unraveling the Taste Fabric of Social Networks

and more stable than ephemeral taste ethoi. Taste neighborhoods of course, overlap with one another, and the discovery and definition of taste neighborhoods seems even prone to the Ptolemaic dilemma—some nodes must be designated as “center of the universe,” and the choice of these centric nodes can greatly affect the resultant neighborhood definition. Two taste neighborhoods with different Ptolemaic centers are shown in Figure 4.

Taste ethos from spreading activation. While the technical details for the discovery process are potentially lengthy, we sketch a conceptual overview of the implementation here. The raw n by n correlation matrix is reviewed as a classic spreading activation network (Collins & Loftus, 1975). That is to say, activation spreads outward from an origin node to all the connected nodes, then from all connected nodes to each of their connected nodes. The obvious observation here is that in our correlation situation, all nodes are connected to a large percentage of the graph, so our graph is super-connected. However, what makes the spreading activation meaningful is that the strength of the spread activation is proportional to the strength of the PMI along any edge in the graph. The energy of the spreading is also inhibited as the number of hops away from the origin grows, according to a per hop discount rate (e.g., 50%). So, spreading with a low tolerance (or, a high threshold for activation), and outward from “Jazz,” “Yoga” (two-hops away) is reachable, but the energy attenuates before the “Football” (also, two-hops away) node can be activated.

Spreading activation outward from an origin node, the result can be likened to that node’s defeasible (default, in the absence of other inputs or biases) taste ethos. This taste ethos is too small when spreading activation is configured with a modest tolerance. On the other hand, if the tolerance is increased too dramatically, the taste ethos will grow in size but its stability will be undermined due to this well-known problem in graph-based inference: beginning at two hops away, reached nodes lose their semantic relevance to the origin node very rapidly. Think of this as the telephone game effect—playing the childhood game of telephone, the first couple of hops are still recognizable, but recognition often rapidly tapers off after the first couple of hops. The effect is also observed by Google in their PageRank algorithm (Brin & Page, 1998) for scoring the salience of web pages by voting. They noted that high-rank pages tended to link to high-quality

Figure 4. Two Ptolemaically-centered taste neighborhoods, computer generated with the follow parameters—a maximum of 50 nodes in each neighborhood, up to the first three instances of any category type are shown. Spatial layout is not implied by the neighborhood; nodes are manually arranged here.
Unraveling the Taste Fabric of Social Networks

pages, and those to other high-quality pages, but after distance=2, reliability tapered off rapidly.

Mathematical morphology. To discover neighborhoods of taste which are larger than particular node-centric ethoi, but which are still nonetheless stable, we borrow two techniques from the field of mathematical morphology (Serra, 1982; Haralick, Sternberg & Zhuang, 1987) and that are widely used in the image processing literature which appropriates them for object segmentation—morphological opening and thresholding. Morphological opening is the mathematical composition of two operators erosion and dilation, in that order. The intuition is that erosion ‘eats away’ at the boundaries of an object, whereas dilation ‘grows’ the boundaries of the object. However, erosion and dilation are not inverses because both erosion and dilation are injective, that is, they are many-to-one and lossful transformations. The effect of morphological opening is also quite intuitive—it removes small objects ‘disturbances’ and opens up gaps when they are located near a boundary. There is morphological opening and there is also morphological closing which is dilation composed with erosion; closing fills in holes and around boundaries more than opening. We employ opening because it is a bit crisper. Opening eliminates blemishes while closing magnifies blemishes. The other technique, thresholding, is frequently used to post-process an opened image. Applying a fixed threshold to an opened image simply turns every pixel above the threshold to 1, and below the threshold to 0.

Erosion and dilation over spread activations. We choose identity nodes as the centric origins for spreading because they are in general more stable places to start from. This follows the rationale that identities are stronger cultural fixtures than a book or a music album, generally speaking. From the identity nodes, we apply a relatively lenient discount, e.g., 0.8, and spread to define a fairly relevant neighborhood. This is repeated over all identity nodes, begetting an ethos for each identity node. Where ethoi overlap, the max of the node’s energy is taken, rather than the sum of the node’s energies. Now, erosion is applied, trimming back the weakest boundary nodes, followed by a dilation, growing the boundary by adding some energy to all nodes connected to the boundary, pushing some of them over the activation threshold and thus growing the mass. In the current implementation, two iterations of opening are performed, though the meaning of this varies widely with the choice of thresholds and other considerations.

In this manner, larger stable masses of nodes, termed taste neighborhoods, are discovered. Thresholding can help us trim a neighborhood to an arbitrary node-size. For visualizations such as InterestMap, neighborhoods comprised of up to thirty nodes seem visually appropriate. We believe that the application of morphological opening and thresholding to a spreading activation network in order to discover larger stable neighborhoods is a novel use, though we do not evaluate this claim within this paper’s scope.

Summary. This section discussed an implementation of weaving the interest fabric out of social networks. Profiles mined from two social network websites were heuristically segmented and normalized according to a heterogeneous ontology assembled together from a variety of data sources. After normalization, correlation analysis learned the affinities between descriptors, and mathematical morphology over the “raw” fabric enabled taste neighborhoods to be discovered and overlaid onto the fabric. Next, we demonstrate the semantic uses of the taste fabric within application contexts.

WHAT IS A TASTE FABRIC GOOD FOR?

As a rich tapestry of interconnected interests and identities, the taste fabric brings us closer to improvisational manipulations of meaning, and affords us at least three semantic functions—the creation of semantically flexible user representa-
tions, cross-domain taste-based recommendation, and the computation of taste-similarity between people. This section explores these three basic semantic functions in the context of a survey of three applications we have developed. InterestMap is a taste-based recommendation system that leverages interactive visualization of neighborhoods to make the recommendation mechanism transparent, thereby enhancing users’ trust perceptions of the recommender. Ambient semantics uses the taste fabric to facilitate social introductions between two strangers, based on their shared taste. IdentityMirror is a digital mirror for identity self-management. Whereas a real mirror shows you what you look like, IdentityMirror shows you who you are. It explores semantically flexible user representations by allowing time, orderliness, and current events in the world to nuance the representation of the viewer.

Figure 5. Two screenshots of the InterestMap interactive visualization. 5a (top) depicts a user browsing neighborhoods of taste visually. 5b (bottom) depicts a user visualizing his own taste ethos by dragging and connecting interesting nodes to the “who am i?” node.
InterestMap

InterestMap (Liu & Maes, 2005a) visualizes the topology of the taste fabric, and in particular it depicts taste cliques, identity hubs, and taste neighborhoods as a navigatable map. As shown in Figure 5a, users can browse InterestMap’s tapestry of neighborhoods, cliques and identity hubs, or, as depicted in Figure 5b, they can interactively build up their own taste ethoi, by searching for and attaching descriptors to a stationary “who am i?” node. The act of connecting a descriptor to the self is deeper than making a mere superficial keyword association since each descriptor is actually something more like a semantic cloud. Once a user has connected several descriptors to his self, those semantic clouds begin to intersect, overlap, and mingle. They begin to imply that other descriptors, which the user has not selected himself, should be within the user’s taste. Hence, the notion of a visual recommendation.

Taste-based recommendation. InterestMap can, given a profile of the user’s interests, recommend in a cross-domain way, books, songs, cuisines, films, television shows, and sports to that user based on taste. The user’s interests are normalized according to aforementioned processes and mapped into the taste fabric. These nodes in the fabric constitute a particular activation configuration that is unique to the user, and the total situation described by this configuration is the fuzzy taste model of the user. To make recommendations, activation is spread outward from this configuration, into the surrounding nodes. Some nodes in the surrounding context will be activated with greater energy because they are more proximal to the taste signature of the starting configuration. The nodes activated with the highest energy constitute the user’s recommendation. Figure 5b shows a visualization of the recommendation process. The user’s self-described interests are the descriptors directly connected to the “who am i?” node. Each of these interests automatically entails other strongly connected descriptors. This is visually expressed well in the InterestMap visualization because a node getting pulled toward “who am i?” will tug a whole web of nodes behind it. Since the visualization starts with just the “who am i?” node visible on the screen, specifying just a couple of interests can literally fill up the screen with its taste entailments. To visualize the spreading activation mechanism, the size and yellowness of nodes diminishes as activation spreads outward from the “who am i?” node.

Visual recommendation enhances transparency and trust. That a user trusts the recommendations served to him by a recommender system is important if the recommender is to be useful and adopted. Among the different facilitators of trust, Wheeless and Grotz (1977) identify transparency as a prominent desirable property. When a human or system agent discloses its assumptions and reasoning process, the recipient of the recommendation is likely to feel less apprehensive toward the agent and recommendation. Also in the spirit of transparency, Herlocker, Konstan, and Riedl (2000) report experimental evidence to suggest that recommenders which provide explanations of its workings experience a great user acceptance rate than otherwise.

Unlike opaque statistical mechanisms like collaborative filtering (Shardanand & Maes, 1995), InterestMap’s mechanism for recommendation can be communicated visually. The idiosyncratic topology of this taste fabric symbolizes the common taste tendencies of a large group of people. For example, in Figure 5a, it is plain to see that “Sonny Rollins” and “Brian Eno” are each straddling two different cliques of different musical genres. The rationale for each recommendation, visually represented as the spreading of flow across the network, is easily intelligible. Thus it may be easier for a user to visually contextualize the reasons for an erroneous recommendation, e.g., “I guess my off-hand taste for Metallica situated me in a group of metal heads who like all this other stuff I hate.”
Although we have not yet implemented such a capability, the ability to interact with the Interest-Map network space would also afford the system an opportunity to learn more intelligently from user feedback about erroneous recommendations. Rather than a user simply stating that she did not like a particular recommendation, she could black out or deprecate particular clusters of the network which she has diagnosed as the cause of the bad recommendation, e.g., “I’ll black out all these taste cliques of heavy metal and this identity hub of ‘Metal Heads’ so the system will not make that mistake again.”

**Ambient Semantics**

Ambient Semantics (Maes et al., 2005) is a wearable contextual information system that supports users in discovering objects and meeting people through pithy just-in-time feedback given in the crucial first moments of an encounter. Here is an example of a use case involving the discovery of a new book: Wearing the Ambient Semantics RFID reader wristband, you pick up a copy of Marvin Minsky’s “Society of Mind” book. Through your cell phone display, the system tells you that you would be particularly interested in section 3 because it is relevant to your current research topics. It would tell you that your friends Henry and Barbara listed this book among their favorites, and that the author’s expressed opinions seem sympathetic to your own, based on semantic analyses of both your writings. The system can indicate that you would find the book tasteful because it can use taste fabric to detect that it is indeed within close proximity to your taste ethos, translating to a strong taste-based recommendation.

**Exposing shared taste-context between two strangers.** The second use case concerns the system facilitating social introductions by breaking the ice. This scenario demonstrates using the taste fabric for the quantification and qualification of the taste-similarity between two strangers. First, a scenario. You are at a business networking event where Ambient Semantics wristwatches have been given to the attendees. You are tired of the same old conversation starters—what’s your name—who do you work for—how do you like it here?—so you head to the Ambient Semantics kiosk where people are meeting each other in a new way. You introduce yourself to a lady standing next to you. By virtue of your handshake, the physical surroundings are transformed. The music and lighting in the area change to suit the shared aspects of yours and the lady’s tastes. Some graphics of kayaking are thrown up on the kiosk display, as well as the faces of some people. The lady says to you, ‘so you know Bob and Terry too? Are you in the Boston Outdoor Society too?’

**Calculating taste-similarity: quantitatively vs. qualitatively.** There is more than one good way to use taste fabric to calculate the taste-similarity of two people. The more direct way is to measure the intersection of two spread activations. Taking each person’s seed profile of interests and mapping it into the taste fabric, we arrive at an initial configuration. Spreading activation outward from this configuration defines a semantic neighborhood, which earlier in the paper we referred to as a person’s taste ethos. Taking the semantic intersection of two or more persons’ ethoi, we arrive at the quantitative calculation of taste-similarity.

However, another intriguing possibility is to make a qualitative calculation about taste-similarity. Although the intersection of two taste ethoi is mathematically satisfying, it is not easily explainable and articulated. In other words, having the system explain that “the two of you share taste because you both have interests x, y, and z in your spreading activation clouds” is inappropriate. More articulate would be to cite a shared habitation of taste neighborhoods, for example, this explanation—“the two of you share taste because both of you are adventurers and lovers of wine.” Here, the mechanism of the recommendation feels more transparent. To calculate qualitative similarity, each person’s
taste ethos would be used to score the degree of a person’s habitation across the various taste neighborhoods, which as you recall, are centered around identity nodes. Like the classic k-nearest neighbors classification scheme, here we classify persons by their k-nearest taste neighborhoods. Having completed this mapping, the subset of neighborhoods shared among the two or more persons become those persons’ shared situation. To communicate shared neighborhoods to the persons, the neighborhoods could be effectively visualized on a screen, or, neighborhoods are safely summarized by stating the identity nodes which live within that neighborhood.

**IdentityMirror**

What if you could look in the mirror and see not just what you look like, but also who you are? Identity mirror (Figure 6) is an augmented evocative object that reifies its metaphors in the workings of an ordinary mirror. When the viewer is distant from the object, a question mark is the only keyword painted over his face. As he approaches to a medium distance, larger font sized identity keywords such as “fitness buffs”, “fashionistas”, and “book lovers” identify him. Approaching further, his favorite book, film, and music genres are seen. Closer yet, his favorite authors, musicians, and filmmakers are known, and finally, standing up close, the songs, movies, and book titles become visible.

The Identity Mirror learns and visualizes a dynamic model of a user’s identity and tastes. Looking into it, the viewer’s face is painted over with identity and keywords, sourced from this dynamic user model. Taste fabric is used to interpret an initial seed profile into a semantic situation within the fabric. For instance, the viewer specifies that he listens to “Kings of Convenience” and enjoys the fiction of Vladimir Nabakov, and using this, taste fabric situates the viewer within its multiple neighborhoods of taste. The keywords which paint over the viewer’s face represent his context within taste-space.

**Dynamic model of taste.** The richness of connections in the taste fabric allow for a truly dynamic model of a user’s taste—one that can evolve over time, and can absorb the influence of each morning’s world events. First, we explain how world events effect the dynamic user model. In Liu (2003), one of the authors gave a model of context-sensitive semantic interpretation for noun phrases. The phrase “fast car,” under a default context, would mean “a car that is moving fast,” but uttered at a carwash, it could also mean “a car that is washed fast.” Similarly, a person’s identity can be interpreted differently based on each morning’s news. For example, supposing that the morning news reveals an international...
conspiracy, that event could tease out from the user’s taste ethos her one-time penchant for Tom Clancy mystery and suspense novels, even though on any other day, that part of her taste would not have been visible. IdentityMirror implements a feature to allow world events to bias who you look like. It operationalizes this by parsing each morning’s news feed for major topics, and activating those topics in the taste fabric as contextual biases. When the user’s taste ethos is generated by spreading activation away from a starting configuration, the activation would now flow in an unusual way because of the new contextual biases. Thus, the image of the user’s identity has been biased by world events.

Second, we explain how the user’s model can evolve over time by recording the history of changes that a viewer makes to himself in the mirror. By gazing into IdentityMirror, a viewer can glean his identity-situation. Is his hair out of place? Are one of his interests out of place? How do his facial features combine to compose a gestalt? How do his various interests come together to compose an identity or aesthetic gestalt? We implement a feature called Identity Fixing which allows a viewer to “fix” himself as he would fix his hair. Keywords are distributed between a hearth (keywords that are taste-cohesive) and a periphery (outlier keywords seemingly out-of-place about a person); the hearth covers the face, the periphery covers the hair. A person with a strong degree of taste-coherence has ruly hair, whereas a postmodernist with scattered interests has unruly hair. The viewer can use his hands to adjust his hair—he can dishevel those unwanted peripheral keywords, or accept them by packing them into his hair. In the user model, rejecting a keyword de-energizes that keyword in the user’s taste ethos, whereas affirming the keyword energizes that keyword in the user’s taste ethos. As world events tease out new and hidden aspects to the viewer’s taste over time, and as the viewer continues to fix his identity, over time, the viewer’s identity will be well groomed and even well vetted.

ADVANCED DISCUSSION

In this section, we present an evaluation of the taste fabric, present related work, and discuss other ways in which this work is of consequence to the semantic mining and Semantic Web communities.

Evaluation

We evaluate the quality of the taste fabric apropo a telos of recommendation, scrutinizing the performance of recommending interests via spreading activation over the taste fabric, as compared with a classic collaborative filtering recommender. Much of this discussion is adapted from (Liu & Maes, 2005a).

In this evaluation, we introduced three controls to assess two particular features: (1) the impact that identity hubs and taste cliques have on the quality of recommendations; and (2) the effect of using spreading activation rather than a simple tally of PMI scores. Notably absent is any evaluation for the quality of the produced taste neighborhoods, because here we consider only quantitative and not qualitative recommendation. Qualitative recommendation is not claimed to outperform quantitative recommendation in terms of accuracy—our suggestion was that linguistically identifying and visually illustrating two persons’ cohabitations of taste neighborhoods should facilitate trust and transparency in the recommender’s process.

In the first control, identity descriptor nodes are simply removed from the network, and spreading activation proceeds as usual. In the second control, identity descriptor nodes are removed, and \( n \)-cliques where \( n > 3 \) are weakened. The third control does not do any spreading activation, but rather, computes a simple tally of the PMI scores generated by each seed profile descriptor for each of the 11,000 or so interest descriptors. We believe that this successfully emulates the mechanism of a typical non-spreading activation
item-item recommender because it works as a pure information-theoretic measure.

We performed five-fold cross validation to determine the accuracy of the taste fabric in recommending interests, versus each of the three control systems. The corpus of 100,000 normalized and metadata-expanded profiles was randomly divided into five segments. One-by-one, each segment was held out as a test corpus and the other four used to train a taste fabric using PMI correlation analysis. The final morphological step of neighborhood discovery is omitted here.

Within each normalized profile in the test corpus, a random half of the descriptors were used as the “situation set” and the remaining half as the “target set.” Each of the four test systems uses the situation set to compute a complete recommendation—a rank-ordered list of all interest descriptors; to test the success of this recommendation, we calculate, for each interest descriptor in the target set, its percentile ranking within the complete recommendation list. As shown in (2), the overall accuracy of a complete recommendation, $a(CR)$, is the arithmetic mean of the percentile ranks generated for each of the $k$ interest descriptors of the target set, $t_i$.

$$a(CR) = \frac{1}{k} \sum_{i=1}^{k} \text{percentile}(t_i, CR) \quad (2)$$

We opted to score the accuracy of a recommendation on a sliding scale, rather than requiring that descriptors of the target set be guessed exactly within $n$ tries because the size of the target set is so small with respect to the space of possible guesses that accuracies will be too low and standard errors too high for a good performance assessment. For the TASTEFABRIC test system and control test systems #1 (Identity OFF) and #2 (Identity OFF and Taste WEAKENED), the spreading activation discount was set to 0.75. The results of five-fold cross validation are reported in Figure 7.

The results demonstrate that on average, the full taste fabric recommended with an accuracy of 0.86. In control #1, removing identity descriptors from the network not only reduced the accuracy to 0.81, but also increased the standard error by 38%. In control #2, removing identity descriptors and weakening cliques further deteriorated accuracy slightly, though insignificantly, to 0.79. When spreading activation was turned off, neither identity hubs nor taste cliques could have had any effect, and we believe that is reflected in the lower accuracy of 73%. However, we point out that since control #3’s standard error has not worsened, its lower accuracy should be due to overall weaker performance across all cases rather than being brought down by exceptionally weak performance in a small number of cases.

We suggest that the results demonstrate the advantage of spreading activation over simple one-step PMI tallies, and the improvements to recommendation yielded by identity and taste influences. Because activation flows more easily and frequently through identity hubs and taste cliques than through the typical interest descriptor node, the organizational properties of identity and taste yield proportionally greater influence on the recommendation process; this of course,
is only possible when spreading activation is employed.

**Related Works**

*A cultural metadata approach to musical taste.* Whitman and Lawrence (2002) developed a metadata model for characterizing the taste coherence of musical genres. Mining adjectival and noun phrases collocated with musical artist discussions in newsgroups and chatrooms, they applied machine learning to automatically annotate music artists with what they termed “community metadata.” Then Whitman and Smaragdis (2002) applied community metadata to build cultural signatures for music genres that could be used, in conjunction with the auditory signal, to classify unknown artists based on style similarity. Their notion of a metadata signature for musical styles is sympathetic to our notion of taste ethos and taste neighborhood, and both systems take a bottom-up metadata-driven view of meaning definition. A chief difference between our two works is that taste knowledge is located in descriptive word-choice in their system (e.g., “wicked,” “loud”), and located in interest-choice in our system, that is, the choices of what people consume (e.g., “Britney Spears”, “Oreo cookies”).

**Social information filtering.** In prior work, one of the authors co-developed a well-known technique for item recommendation based upon nearest taste-neighbor, the approach known variously as social filtering, or collaborative filtering. Shardanand and Maes (1995) represent users as vectors of (item, rating) pairs, and compute taste-similarity as statistical correlation between user vectors, or alternatively as cosine similarity of vectors in n-dimensional item space. In their Ringo social music browsing system, users were recommended a list of potential ‘tastemates’ on the basis of taste-similarity. One difference between our two approaches is that social filtering maintains distinct user profiles, whereas taste fabrics dissolves user boundaries, and is, in their terminology, a ‘content-based filtering’ approach. In distilling a reusable knowledge resource out of social network profiles that can be reappropriated for a variety of other purposes not concerned with the original social network community, it is necessary to protect the privacy of the original users, and we suggest that taste fabrics serves as a model for doing so. Also relevant is Sarwar et al.’s (2001) item-based collaborative filtering approach to recommendation, which, like taste fabrics, relies upon item-item correlation rather than user-user correlation. Taste fabric exceeds item-based filtering by use of extensive metadata to ‘relax’ the meaning from the item itself, by defining identity descriptors as supernodes, and by representing users as k-nearest neighborhoods. In general, collaborative filtering is more representation-opaque whereas spreading activation over neighborhoods can be visualized and more easily debugged.

**Social network analysis and relational mining.** Much research has examined the explicit structure of social networks, and studied their topologies via graph theory. Newman (2001) mined scientific coauthorship networks and found that collaborations ‘funneled’ through gatekeeper scientists. In taste fabrics, identity hubs, and hubs created around particularly salient interest descriptors constitute a similar topological feature. Jensen and Neville (2002) mined structured metadata relations from the Internet Movie Database (imdb.com) called ‘schema’ and learned a Bayesian network model to represent and predict item distances probabilistically. They also model the relational semantics of social network relations implied between movie actors from the Internet Movie Database and the Hollywood Stock Exchange (www.hsx.com). Finin et al. (2005) examine how the FOAF (“friend-of-a-friend”) ontology applies Semantic Web concepts to enable efficient exchange of and search over social information, illustrating how social networks could develop with its semantics already explicit. Finally one work which considers the semantic
content entailments of social network users is McCallum, Corrada-Emmanuel, and Wang’s (2005) modeling of Author-Recipient-Topic correlations in a social network messaging system. Given the topic distributions of email conversations, the ART model could predict the role-relationships of author and recipient. The work considers group clusters and dyadic relationship dynamics but does not consider cultural aggregates as is the concern of our present work.

**Large-scale commonsense knowledge networks.** Taste fabrics are a rich tapestry which define the meaning space of taste and interests. They are represented as semantic networks and reasoning is performed via spreading activation over this network. This approach to knowledge representation and reasoning builds upon previous work in large-scale semantic knowledge bases such as WordNet (Fellbaum, 1998) and ConceptNet (Liu & Singh, 2004). WordNet is a semantic network whose nodes are words, and edges are various nymic lexical relations between the words, e.g. a “dog” has the hypernym of “canine.” ConceptNet is a semantic network of commonsense knowledge whose 200,000 nodes are verb phrases (“eat burger”, “take shower”), and 1.6 million edges are one of 20 kinds of world semantic relations (e.g., “EffectOf,” “PropertyOf,” “DesireOf”), e.g., (EffectOf “be hungry” “cook food”). ConceptNet and taste fabrics reason similarly by activating a seed configuration of nodes, and spreading activation outward to define a semantic context. Both resources are densely connected, semantically extensive within their respective domains, and allow for improvisational manipulations of meaning to take place atop them.

**Reusable Methodologies**

**Sanitary semantic mining.** The sanitariness of a mined knowledge resource is the degree to which it is purged of idiosyncrasy, especially idiosyncratic traces of user-specific information, and also idiosyncrasies which implicate the original application domain from which the resource was mined. When a knowledge resource is sanitary, assurances can be made that private user data is not recoverable, and that the resource is sufficiently context-free so that it could potentially be used to solve problems across a variety of domains. Taste fabrics are an illustration of how a sanitary knowledge resource can be mined out of a highly idiosyncratic and application-specific data source such as self-descriptive social network profiles. Because it is sanitized, taste fabrics can be publicly distributed and used to power applications living in other domains.

When mining social network data, concern for privacy and copyrights of user data make derivative works especially problematic; yet there is a great need and opportunity to infer valuable semantic knowledge from these sources. Ensuring data anonymity in the produced knowledge resource is a particularly sensitive issue. An early phase of the taste fabric construction process is to normalize the casually-stated keywords and phrases into formal ontologies of non-idiosyncratic form (e.g., “Nietzsche” into “Friedrich Nietzsche”, “dogs” appearing under the “passions” category into “Dog Lover”). Already, the unrestricted idiosyncratic language which bears traces of an authorship are beginning to be wiped away. In contrast, collaborative filtering systems maintain ratings for each user, and while users do not have to be named, even unnamed users are not anonymous, they are only pseudonymous. A user’s name is simply wiped away and replaced with a unique id (renamed from “John Smith” to “User #123”), but the profile’s integrity is intact. Because the number of instances is quite large in the space of tastes, it may be possible to recover the identities of pseudonymized users because the constitution of profiles are quite unique. At the very least, maintaining any information structured around the notion of a user lends itself to the perception that privacy of the source data may be violated.
Rather than preserving individual profiles, the taste fabric simply uses these profiles to learn the strengths of connections on a network whose nodes already exist (they are simply an exhaustively enumeration of all features in the ontology). The method of the learning is non-linear so explicit frequency counts cannot easily be recovered. Thresholding and neighborhood definition are further lossful transformations which make details of the original application data virtually unrecoverable. The final structure is sanitary—it assures the anonymity of the data source, and is much easier to distribute.

Instance-based semantic webs and ethotic representation. In the Semantic Web community, ontology and metadata systems are often seen as top-down and bottom-up approaches to knowledge representation, respectively. To draw parallels with the artificial intelligence literature, ontology is a category-based representation, and metadata is a feature-based representation. However, taste fabrics introduces the notion of an instance-based representation, which we feel to be a promising methodology for the Semantic Web community that warrants further study, especially into the issue of scalability. An instance-based representation lacks categories or features, having only items and dense numerical connections between them. Knowledge is thus unpacked from the linguistic symbolism of a category or feature’s name, and instead, is found in connectionism—the flow of semantics through a graph of items. The shift from symbolic interpretation toward continuous interpretation parallels Zadeh’s efforts in attempting to soften the bivalence of logic representation by giving a fuzzier, more continuous account of meaning (Zadeh, 2004).

Instance-based representations are more appropriate for semantic recognition and semantic mediation because they offer continuous numerical interpretation of entity similarity. In taste fabrics, users, groups of users, and cultures can all be represented uniformly as clouds of node activations in the fabric. A taste fabric allows the meaning of a user’s keyword profile to be ‘relaxed’ into a semantic cloud which we term an ethos. Using ethotic representation, semantic mediation between two users or entities in the fabric can be computed quite easily as shared activation, and even effectively visualized. By interpreting an ethos as a membership into k-neighborhoods, the resource can be used to classify users or entities into an ontology of neighborhoods (the organizing force of ontology, in fact, is still present in the resource via neighborhoods and identity descriptors). Instance-based representations and ethotic representations would be well-suited for semantic resources meant for mediation and classification in the Semantic Web.

CONCLUSION

This paper presented a theory and implementation of taste fabrics—a semantic mining approach to the modeling and computation of personal tastes for lifestyle, books, music, film, sports, foods, and television. Premised on philosophical and sociological theories of taste and identity, 100,000 social network profiles were mined, ontologically-sanitized, and a semantic fabric of taste was weaved. The taste fabric affords a highly flexible representation of a user in taste-space, enabling a keyword-based profile to be ‘relaxed’ into a spreading activation pattern on the taste fabric, which we termed a taste ethos. Ethotic representation makes possible many improvisational manipulations of meaning, for example, the taste-similarity of two people can be computed as the shared activation between two ethoi. Taste-based recommendation is already implied by a taste ethos, as all items within an ethos are intrinsically relevant to the taste of the individual. Indeed, an evaluation of recommendation using the taste fabric implementation shows that it compares favorably to classic collaborative filtering recommendation methods, and whereas collaborative filtering is a opaque mechanism,
recommendation using taste fabrics can be effectively visualized, thus enhancing transparency and cultivating user trust.

Two models of taste-based recommendation—one quantitative based on shared activation, and one qualitative based on k-nearest neighborhoodswere presented. Recommendation, time and world-sensitive user representation, and interpersonal taste-similarity, were illustrated within a survey of three applications of taste fabrics.

This paper makes three contributions to the literature. First, it presents a novel mechanism for mining and modeling the taste-space of personal identities and interests. Second, the mining and weaving of taste fabrics from idiosyncratic social network profiles raises the issue of sanitation of knowledge resources, and this paper illustrated how ontology and non-linear correlation learning can be used to purge idiosyncrasy and prepare a general-purpose grade knowledge resource. Finally and third, in addition to ontology-based and metadata-based knowledge resources, taste fabrics introduces a novel third approach to the literature—instance-based fabrics, where the notion of ‘knowledge’ is a purely relational one. Fabrics, we suggest, excel at semantic mediation, contextualization, and classification, and may play a valuable role as a context mediator in a recently complicated Semantic Web of formal, semi-formal, and now, informal, entities.

ACKNOWLEDGMENTS

This research was supported by a British Telecom Fellowship, an AOL Fellowship, and by the research consortia sponsors of the MIT Media Lab.

REFERENCES


Herlocker, J., Konstan J., & Riedl, J. (2000). Ex-
plaining collaborative filtering recommendations. *Conference on Computer-Supported Cooperative Work* (pp. 241-250).


**ENDNOTES**

1. http://www.dmoz.org
9. A qualifying clique edge is defined here as an edge whose strength is in the 80th percentile, or greater, of all edges.
10. By discounting a random 50% subset of the clique’s edges by a Gaussian factor (0.5 mu, 0.2 sigma).

*This work was previously published in International Journal on Semantic Web & Information Systems, Vol. 2, Issue 1, edited by A. Sheth & M. Lytras, pp. 42-71, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 4.8

Computer-Mediated Communication: Enhancing Online Group Interactions

J. Michael Blocher

Northern Arizona University, USA

ABSTRACT

Advancements in information technology have transformed much in all aspects of today’s society. In particular, synchronous and asynchronous electronic communication systems support countless interactions every day. However, these communication systems have specific limitations that can cause miscommunications as well. Because of that, computer-mediated communications (CMC) has been a robust research agenda in many disciplines. This is especially true of education, where online learning has become common place in all but a few universities, thus requiring learners to interact via CMC within virtual learning environments. This chapter will use educational CMC research as a lens to promote and support an understanding of how to better utilize and facilitate electronic communication, regardless of the field or endeavor.

BACKGROUND

Over a decade ago, Rheingold (1993) described how the Internet, specifically, synchronous and asynchronous electronic communication could create a non-linear and level environment that provides the conduit for human interaction that is culturally neutral—where members meet in virtual communities judged by ideas, thoughts, and contributions, rather than by race, gender, age, physical appearance, or national origin. In particular, CMC was described as a venue where participants could engage in discussions on bulletin boards and listservs with equal impunity and have the opportunity to create a cyber community. Due to the lack of visible and contextual cues, participants engaging in this, somewhat, anonymous form of communication were free from prejudice and able to speak their mind more freely and openly than in traditional face-to-face
group interactions. However, this turned out be only somewhat correct, and the literature provides ample examples of how CMC group interactions follow similar patterns and styles that have been identified within traditional face-to-face group interactions.

Communication Style Differences

Communication style differences, particularly, gender-based differences have been identified as one area that has been investigated in more historic group interaction literature. Specifically, women tend to engage in group interactions with a style distinctly different than men. Eakins and Eakins (1978) in researching gender styles within group interactions discovered that women tend to be more engaged in developing intimacy and maintaining relationships than their male counterparts. In other words, women tend to be more interested in building relationships and developing closeness within their group than men do. For example, women are more apt to ask for another’s opinion, desiring to engage others within the conversation. Men, on the other hand, tend to be more directive and less apt to draw out the interaction (Fishman, 1983). Perhaps this is because, in general, women are more interested in the social aspects of group interaction. For example, Carli (1984) reported that women are more social-oriented, while men are more task-oriented in their interactions. Furthermore, Briton and Hall (1995) reported that women tended to be better at using nonverbal cues, being more expressive, utilizing body language, eye contact, and gestures to send and receive subtle messages to promote and enhance communication in group interactions. Men, on the other hand, leaned toward competitiveness and dominance seemed more apt to be more competitive and less socially motivated in their group interactions—using fewer subtle nonverbal cues.

In terms of group interactions where productivity is at issue, Maguire (1999) reported that within studies of marketing executives women work more to build consensus while men tended to make faster decisions. One could argue that one style is better than the other at supporting communication in group interactions, as both have benefits and limitations. For example, being more direct could be seen as a way to improve communication by not clouding the issue with possibly vague nonverbal cues. On the other hand, by providing more information via other channels (e.g., body language, facial expressions, etc.), the verbal message might be further enhanced, with less chance of misunderstandings. Furthermore, by utilizing a communication style that seems competitive and dominating, issues of power may cloud or hinder the real message. Again, on the contrary, using less direct verbal communication could be seen to hinder one’s ability to provide leadership. Which style makes a better communicator—one who builds community or consensus through social group interactions enhanced with non-verbal cues, or one who improves productivity with more directive style communication?

It could be argued that it might be better to try to understand these communication style differences, rather than trying to debate which might be more effective, as each has benefits and limitations. This is especially important within the CMC environment. As we will see in the next section, CMC group interactions have other issues based upon the various CMC media, which can greatly impact these more traditional communication style differences.

CMC Communication Style Differences

Of course, these gender differences are generalities. One’s individual group communication style is just that, individual. However, these gender-related communication style differences seem to parallel a CMC environment as well. As in all social environments, membership within a CMC group environment can be dominated by
more spirited individuals, and certain hierarchies can develop or emerge based upon the members’ interactions within that virtual community. Of particular note, Herring (1993) details that virtual communication spaces may not be the democratic spaces described earlier—primarily because of the communication style differences brought naturally to the environment by the various group members and the community ethos that the group develops. For example, her study reported that men generally made “strong assertions,” engaged in “self-promotion,” asked “rhetorical questions,” came from an “authoritative orientation,” provided “challenges and humor.” Women, on the other hand, tended to ask “questions,” provided “attenuated assertions,” had a “personal orientation” and provided “support” in their CMC. As discussed above, these two gender-based communication styles mirror those of the traditional research. Generally, women tend to foster consensus and community building, while men tend to demonstrate a communication style that is more directive in nature.

While this supports the argument that men are more apt to utilize a more directive CMC style, perhaps more interesting are specific investigations into chat (synchronous communications) that have detailed how gender communication style differences were power-based in particular. Soukup (1999) stated that in his study, “Masculine participants were aggressive, argumentative, and power oriented. While feminine participants sought relationships and intimacy, they were often dominated and overpowered by the aggressive discourse of the masculine members (p. 169).” Perhaps, even more disturbing, findings were reported by Herring (1999). She described male-dominated harassment of female participants in two case studies, stating: “Due perhaps to the youth and sexual preoccupations of the majority of its users, Internet relay chat (IRC) tends to sexualize female participants, and to involve them in flirtatious interactions. Accordingly, the ideal for female interactions in IRC appears to be cooperative flirtation, as it is also in many off-line recreational social realms” (p. 163).

One would think that this type of interaction might inhibit female members when engaging in online communication, perhaps making them more apt to adjust or change their online communication engagement; perhaps even hide or mask their gender identity in CMC environments to guard against this type of interaction. To answer that question, Jaffe, Lee, Huang, and Oshagan (1999) investigated how participants, when given opportunities, chose to select anonymous pseudonyms for their CMC. Some participants chose pseudonyms with same gender, others chose pseudonyms that were cross gender, and finally some chose pseudonyms that were gender neutral. Gender-neutral and cross-gender pseudonyms then collapsed into gender-masking, thus providing two categories as same-gender and gender-masking. They reported that, “Seventeen (81%) of the 21 male participants in the pseudonymous conference chose pseudonyms that retained their gender identification. In contrast, 13 (81%) of the 16 females in the pseudonymous conference chose pseudonyms that masked their gender identification” (1999, p. 227). This significant difference (p < .001) between men and women demonstrates a greater tendency, by women, to utilize a pseudonym that masked their gender. The authors of this study saw this as evidence of a direct strategy to level the playing field stating:

*This finding is interpreted to reflect an effort to rectify an imbalance, felt by women, of social parity when interacting in mixed-gender situations. An especially pertinent example of previous gender-based naming practices provides historical justification for such sentiments. In anecdotal recollections within the scientific research profession, female authors were often referred to by first and last names, while only initials and last name identify males.* (Jaffe, Lee, Huang, & Oshagan, 1999, p. 230)
The CMC literature provides examples where group interactions within CMC environments seem to exacerbate the communication style differences, perhaps making men aggressive and making women more engaged in building social relationships in their communication styles. However, it seems that if given the opportunity, women tend to mask their gender for strategic purposes.

**Summary**

Although the CMC technology does have the potential to provide a non-linear and level environment that provides the conduit for human interaction that is culturally neutral, it seems that it may not. Indeed, the studies described above strongly indicate that CMC may not be the level playing field that it was once thought to be. In particular, there are communication style differences within the CMC environment that parallel traditional communication styles, thus creating a group communication environment where the communicators interact in similar ways that they would in traditional face-to-face environments—where gender, power, and communication style may have a strong impact on one’s communication within group interactions.

It is important to remember that these are general tendencies and that not all CMC group interactions are destined to exhibit the type of interactions described above. Rather these are tendencies that could impact one’s experience engaging in CMC, which might be exacerbated by some of the unique communication issues that are specific to current text-based CMC. For example, often a message is sent that is totally benign. However, the message even when taken simply for the written text may be taken for a completely different meaning than originally intended, because the sender and recipient engaged in the communication at different times, locations, and states of mind. Being more knowledgeable about the tendencies and limitations of CMC could help limit more adversarial interactions that were based upon miscommunication due to CMC issues or communication styles. In particular, becoming aware of and monitoring one’s personal communication style could help make the individual more thoughtful and strategic about their communiqués, thus reducing the tendencies for miscommunication to occur. The next section will discuss specific issues at work within the current CMC environments that might hamper communication and strategies that could be implemented to enhance more equitable and clear CMC engagement.

**ISSUES**

As described, gender, power, and communication style can impact one’s personal communication interactions within a CMC environment. Although these elements parallel traditional face-to-face communication interactions, there are issues with CMC that are unique to that environment. In particular, the discussion will focus on current CMC issues in terms of how asynchronous and synchronous electronic communication tools limit the capacity of communication cues and meta-information as well as communication styles in terms of gender, and whether these issues can be influenced to help ameliorate those limitations.

**Media Richness**

Much of the CMC literature suggests that communicating within electronic messaging systems, which are primarily text based, limits communication because of the lack of other communication cues or components. Media richness theory (Daft & Lengel, 1986) defines communication media in terms of face-to-face communication elements. These elements include immediacy of feedback, non-verbal, and other cues that enrich communication. In other words, a medium is considered richer or thicker when it can sup-
port a greater number of cues to communicate a person’s ideas (e.g., facial expressions, gestures, voice inflection, etc.). In light of media richness theory, current CMC is defined as being “lean” because much is asynchronous communication, lacks immediacy of feedback and generally lacks “multiplicity of cues.”

In terms of immediacy of feedback, imagine a face-to-face interaction where participants give and receive concurrent and instant feedback while they monitor the message reception and consider its impact. Compare that to asynchronous CMC where interactions are often delayed by hours, if not days. Indeed, can a sender remember the exact meaning of a message or state of mind they were in upon sending a message when they receive the reply a day or two later? Even synchronous CMC (chat) lacks in immediacy, where there is a lag time of waiting for the individual to complete their typing. Furthermore, synchronous chat is dependent upon the individual’s keyboarding skills.

In addition to the limitations of asynchronous communication, text-based CMC also does not include non-verbal cues, inflection, and other elements of face-to-face interactions. As most current CMC messages are text only, they are devoid of very rich communication cues that can cause miscommunication. These missing cues can impact CMC with possible miscommunications in several ways. For example, often in verbal communication of most languages inflection describes how the message is to be taken. English speakers often use inflection to relay whether something said is a statement or a question. Punctuation provides cues in the written language to account for this issue. However, many times in CMC group interactions, communicators use less than formal writing styles, which often can cause confusion. An additional example is sarcasm. While one might fully understand one’s intent of a sarcastic statement in a verbal interaction, the message can very easily be mistaken in a written interaction.

**Media Richness’ Impact on Communication Style**

Can media richness in CMC impact one’s ability to communicate, depending upon an individual’s communication style? In other words, if one’s communication style utilizes communication cues that are missing in the leaner CMC media, will it create a less equal communication environment? To help answer that question Dennis, Kinney, and Hung (1999) investigated decision-making task performance in terms of media richness and gender differences. In particular, they attempted to answer questions of the impact of media richness on performance in terms of “decision time,” “decision quality,” “consensus change,” and “communication satisfaction.” The study investigated gender-based teams performing equivocal decision-making tasks both in synchronous chat and face-to-face. As women tend to better utilize and read non-verbal cues, which are missing in CMC, they were especially interested in the results of the impact of gender on the performance tasks. Teams of all male, all female, and mixed gender were formed and compared performing equivocal decision-making tasks in both face-to-face and chat environments. The results indicated that only “decision time” was impacted by the use of CMC. Indeed, “decision quality,” “consensus change,” and “communication satisfaction” were not significantly impacted by the use of CMC. Interestingly, the all-female team took five times as long to complete an equivocal task of reaching a decision utilizing CMC than it did face-to-face. This supports the idea that CMC has a more profound impact on women than men due to the lack of cues that women tend to better utilize as they needed more time to complete the tasks. More interesting, however, is that the decision quality of the all-female team utilizing CMC was higher than the all-male and mixed-gender team; second only to the all-female face-to-face decision-making team. In other words, while missing the nonverbal cues, it took this team much longer
to perform the task, but the female members still took the time to perform at a high level regardless of media—suggesting that they could, and did, modify their engagement to successfully complete the task.

This study provides evidence that media richness, or leanness of, in this case, does impact one’s ability to engage in CMC group interactions. More specifically, since women tend to utilize nonverbal communication cues more often, a CMC medium that strips out those communication cues, can provide a distinct gender disadvantage. However, it is very important to remember here that these are tendencies. Indeed, anyone who utilizes a communication style that is supported by nonverbal communication cues and engages within a CMC group interaction where those cues are missing due to the medium’s leanness is at a disadvantage—regardless of gender.

**CMC Strategies**

From the study detailed above it seems that the participants modified or enhanced their communication within the limited media in some way to better perform their tasks. While it is clear that various CMC media have limitations, users can and do use strategies that can help ameliorate the limitations of a particular media. There are specific examples of communicators modifying their communication to better fit a medium. As a case in point, text shortcuts, such as: btw = by the way, irl = in real life, cu ltr = see you later, and imho = in my humble opinion, have been developed and utilized within in CMC environments providing evidence of communicators utilizing new methods to speed up and facilitate their interaction. The practice of abbreviated-text messaging and instant messaging has become common place. Indeed, mobile phone text messaging has become a social phenomenon practiced by young people in many parts of the world. For example, Grinter and Eldridge, (2003) describe how British teenagers are similar to teenagers in other parts of the world in their use of a meta-language of shortened words to send text messages for a variety of purposes, including coordinating future communications within other media, generally phone calls. This trend suggests that these young communicators have developed an understanding and the strategic utilization of various media for a variety but connected purposes.

The above example shows modifying a simple communication method to be more strategically purposeful of the specific communication media. However, as discussed above, gender-based communication style is much more complex. Will one utilize CMC strategies to reduce the limitations of CMC media? Fahy (2003) investigated supportive CMC strategies within an online learning environment to see if participants utilized strategies that would increase interaction and reduce interpersonal distance within the leaner CMC media. Specific strategies were defined as supportive and were seen to replace or approximate “sociolinguistic devices and strategies, including nonverbal elements normally available in face-to-face interaction, (Ridley & Avery, 1979)” (Fahy, 2003. pg. 5). The results suggest that these participants were able to and took the time to utilize important CMC strategies. As Fahy states: “One-quarter (25.9 per cent) of all interaction in these conferences was classified as supportive, a proportion which, though it must be tested against wider and larger conferencing samples, indicates the importance attached to the interpersonal in these online interactions” (Fahy, 2003, p. 11). Students who engaged in these online conferences did, in fact, utilize the CMC strategies that approximated the missing cues, thus enriching the lean medium that promoted a more supportive group interaction.

**Gender-Based Communication Style Identification**

While the CMC literature suggests that men and women tend to utilize different communication...
styles in CMC, one might ask: How strong are these tendencies and can the communication style give away the author’s gender identity in an anonymous interactive environment? Can one judge the gender of a CMC author based on the CMC style without knowing the author or his/her gender? Does one’s awareness of CMC styles provide a tool to distinctly judge an author’s gender? In other words, is a person’s gender given away by his or her communication style even in an anonymous CMC environment? Savicki, Kelley, and Oesterreich (1999) detail a series of studies where participants were grouped in male-only, female-only and mixed-gender groups for CMC interactions. Their anonymous messages were then provided to participants to be judged as to the author’s gender based upon the communication style analysis. They analyzed messages that included a greater number of “I” statements, statements of personal opinion, self-disclosure, and coalition language. “Since this pattern of communication was related to high levels of satisfaction with the group process and high levels of group development, it was labeled the high group development communication style (HCS) and contrasted with the opposite style, low group development communication style (LCS)” (p. 187).

As might be predicted, participants were quite accurate in judging a LCS message authored by men (.708), while significantly less accurate (p < 0.001) in judging an HCS message authored by men (.395). In opposition to prediction, however, participants were significantly less accurate (p < 0.05) in judging HCS messages authored by women (.549) than LCS messages authored by women (.662). These results support the finding that participants were significantly more accurate when judging LCS messages than HCS messages, regardless of the gender of the author. In this study, if a message exhibited HCS it was more difficult to accurately judge the author’s gender than messages that exhibited LCS. They also reported that the judge’s gender or CMC experience did not seem to factor into the level of accuracy as no significant differences for these variables were found.

These mixed results suggest that, while gender identification may be possible, it is quite limited. In other words, if authors utilize HCS within their anonymous CMC messages, there is less likelihood of being identified as male or female. If one’s communication style includes HCS in the CMC messages, they may well be taken for their information and not for their gender. This may have an impact on helping promote a more democratic communication environment based on gender inequities previously discussed.

**Gender-Based Communication Style Training**

Certainly, promoting more democratic CMC group interaction seems a prudent thing to try to achieve. However, can one be trained, and more importantly to become motivated, to utilize HCS to support participation, satisfaction, and group development within a CMC environment? Savicki, Kelley, and Ammon (2002) investigated that question by grouping participants in male only (MO), female only (FO), and mixed gender (MIX). They examined how HCS training would impact participation, satisfaction, and group development. The results were mixed. For example, while they reported that the FO groups scored higher on group development, it may have been due to the potential of gender-based communication style tendencies for that group of females rather than HCS training. However, HCS training for the MIX groups had a greater impact than the MO. The authors state that:

. . . the MIX groups responded somewhat more to the training than the MO groups. The combination of men and women in the same group seemed to change the norms toward more acceptance of HCS communication. Gender is an important variable influence communication in
online groups (Savicki & Kelly, 2000). Gender communication norms seem to be quite strong and should be taken into account when selecting participants for virtual teams. (p. 267)

It seems logical to presume, and from the findings of this study, that it is possible to provide HCS training to impact user engagement. The MIX groups did provide some evidence of the success of the training. However, the HCS training MO groups had less success, according to the authors, because of previously established gender norms. Training individuals to change their communication style for group interactions, in terms of developing higher group development communication skills, may be more complex than first thought. For example, it may not be as simple as providing HCS training, as more social communication interaction norms are at play. However, this study also instigates a question. What interactions in the MIX groups promoted consensus or community building to change the communication norms of accepting increasing use of HCS communication more than the MO groups?

**Summary**

The studies described above detail some of the major issues of CMC. Specifically, current CMC systems tend to be lean media in terms of media richness theory, which can impact certain populations based upon their communication strengths. However, it also would seem that communicators believe it is important and take steps to ameliorate the limitations that the lean media present by utilizing communication methods and strategies that are supportive in nature. Interestingly, utilizing higher group development communication style (HCS) might limit the identification of the communicator, thus possibly providing a more level communication venue in terms of gender. More importantly, however, it seems that the studies above demonstrate that communication styles, although tend to be gender-based, may be a matter of choice. Furthermore, although it seems possible to provide training to support more HCS communication, the process may be quite complex. Indeed, higher group communication development may be greatly impacted by the membership of that group. Although this might seem logical and somewhat simplistic, it could have very complex implications. For example, in becoming knowledgeable about, and strategic in the use of their communication style, CMC users could become better more effective and thoughtful communicators by utilizing a style that is more engaging, supportive, and more likely to foster group development. However, they may not be motivated, or perhaps do not believe they have the permission to do so based upon the membership and communication norms of their virtual group.

**RECOMMENDATIONS**

From the review of the literature there are several issues that dominate the various common CMC systems and the participants who engage in group interactions within them. In particular, there are issues of media richness and communication style differences. In addition, there is the issue of a group’s communication norms. These factors can impact the group’s communication norms, which might be defined by: (A) the makeup of the group membership, (B) the purpose of the group, and (C) the media they employ to communicate. These issues can provide some challenges for CMC users. Drawing from distance or distributed learning theory will provide some recommendations to help support electronic communication by building online communities where group interaction is enhanced by developing membership within a community of practice. In addition, the literature suggests CMC strategies that might help communicators better their electronic communiqués, thus reducing miscommunications.
and more equitable group interactions. Finally, current CMC technologies generally provide for text-only communication. In particular, there is the issue of media richness with communication environments that utilize a suite of tools that include only the traditional text-based synchronous and asynchronous e-communication systems. Utilizing multimedia CMC systems that include audio and/or video in addition to the lean text-based synchronous and asynchronous systems can help thicken the communication media, thus alleviating some of the limitations and issues discussed above. All of these will be discussed in the next section.

**Community Building**

As discussed earlier in this chapter, Rheingold (1993) described virtual communities where members interact using CMC within a level social environment. Although this ideal may have been suggested prior to having a greater understanding of some of the media limitations and how group interactions are impacted by them, it does not mean that the idea should necessarily be abandoned. This is especially important as group communication norms can dictate the membership engagement of communication style. One only needs to look at the incredible growth in online tools like, MySpace (n.d.), Orkut (n.d.), and the plethora of blogs to see that people want to communicate and engage in CMC group interactions. For example, more advanced robust open-community CMC systems, such as BuddySpace, are being developed to provide the users with greater flexibility, in terms of connectivity with other platforms. In addition, BuddySpace provides the users’ location, which can help group individuals who might have similar interests because of their locale. However, as with any community, online community members may not be as supportive to one another as they ultimately could, or perhaps should. Therefore, it would be wise to understand the elements of community building as it may impact CMC group interactions in online communities.

In a very well-established text, Kim (2000) provides details and recommendations of community-building strategies that are supported by three underlying design principles: (a) design for growth and change, (b) create and maintain feedback loops, and (c) empower members over time. These principles suggest that the placement of responsibility of community development begins with the designers or developers rather than members. However, it supports the cultivation of member ownership by its three design principles. For example, designing for growth and change, allows for the community to be molded by the feedback received from the members as they gain empowerment. With the empowerment of the members over time, the organizational ownership and responsibility of the community will be cultivated and be accepted by its members. That it not to say, however, that in an online community, the members cannot be the designers and or developers. Regardless, in Kim’s recommendations the members will assume responsibility and ownership of the community.

Another example of community building is detailed in the work of Wenger (1998). In defining a “community of practice,” he provides a detailed outline that identifies a more complex view of communities, stating:

A community of practice defines itself along three dimensions:

- **What it is about:** It’s joint enterprise as understood and continually renegotiated by its members.
- **How it functions:** Mutual engagement that binds members together into a social entity.
- **What capability it has produced:** The shared repertoire of communal resources (routines, sensibilities, artifacts, vocabulary, styles, etc.) that members have developed over time. (p. 54)
Wenger’s description of a community of practice suggests that “members bind together in a social entity” by sharing routines, sensibilities, artifacts, vocabulary, and style. Wenger goes on to say, “Even when a community’s actions conform to an external mandate, it is the community—not the mandate—that produces the practice. In this sense, communities of practice are fundamentally self-organizing systems (p. 2).” The concepts outlined in Wenger’s community of practice have very interesting implications in light of the HCS training findings of Savicki et al. (2002). In that study, the mixed-gender groups moved toward building a community of practice by changing their norms to accept and use more HCS in their CMC within their group interactions. However, one could also argue that the male-only group members also moved toward building a community of practice by not changing their group’s communication norms to accept and use the HCS communication strategies. The HCS training had little or no effect on the male-only group. On the contrary, the group maintained their communication style regardless of the training, perhaps because of the strong influence of the male group communication norms. The external mandate, as Wenger describes, could be seen as the training to change to HCS communication, which was not accepted by the male-only groups, but was more so by the mixed-gender groups. In this case, the CMC norms were developed and determined by the group. This supports the idea, in terms of online communities, that the ownership and responsibility for the practice of the community lays within the members of the community, not the administrators or developers of an online space.

To provide an example in practice, let us look at designing and developing an online course of study, which utilizes CMC systems to build communities of practice that support enhanced group interactions. It is important to remember that online course delivery has its roots in the distance learning model originally developed that utilized the traditional mail system to send materials to learners in remote locations in the early part of the 20th century. As technology developed, course material delivery moved to more advanced systems: radio, television, CD-ROM, and, eventually, to the Internet. As the delivery methods advanced, opportunities for learner interaction became greater because of the possibilities for group interactions using CMC. Distance learners were no longer limited to primarily one-way information dissemination for their learning experience.

Distributed learning is a term that has been used to replace distance learning by some. Dede (2004) defines distributed learning communities where learners engage collaboratively to develop knowledge as they interact within the various electronic communication and information delivery systems. In this concept of community building, group interactions are distributed via CMC and other more advanced electronic tools, to help members actively engage with one another for the specific purpose of learning. Here, learners within the distributed learning community utilize a variety of resources in a highly structured learning environment. Learning management systems, such as WebCT (n.d.) and BlackBoard (n.d.) include the tools to support this definition of distributed learning.

Although the use of the CMC systems will provide the opportunity for group interactions within a distributed learning environment, the members within the community are still responsible for their interaction and, thus, for their learning. Therefore, designing and developing an online course of study that utilizes distributed learning elements is more complex than using a system to make course materials and opportunities for group interaction available. The key to a distributed learning environment, accordingly, is building communities of practice through the utilization of multiple electronic communication systems that advances the opportunities for the
Computer-Mediated Communication

members to engage, thus making CMC the vital means of group interaction.

In this example, the designer, developer, and/or the instructor begin by selecting the interactive communication systems to utilize. They also will need to design learning activities and assignments that will have the learners engage with the course materials and activities. Finally, they also provide guidelines for group interactions the learners will have with the instructor and their peers. Palloff and Pratt (1999) described effective strategies for building and the cultivation of an online learning community that fosters greater group interaction that augments online learning. They outline a framework for online instruction that provides for: (a) mutually negotiated guidelines, (b) focused outcomes, (c) active creation of knowledge and meaning, and (d) collaborative learning. Interestingly, these elements echo the communities of practice work of Wenger (1998) in terms of sharing the responsibility of community building and, in this case, learning throughout the membership of the community. All of the members, including learners and instructors, become responsible for developing the community of practice. For example, by promoting mutually negotiated guidelines, learners are given the opportunity to share ownership of the learning environment that will bind the members together in a joint enterprise. Learners utilize CMC to communicate with their peers and instructors to mutually produce the shared repertoire of communal resources and learning guidelines that will structure their social entity—the learning community. Through their collaborations they actively create knowledge and meaning based on the focused outcomes, as defined by the community. Designing an online learning environment with these four strategies will then support the members in building a community of practice for learning that utilizes the pedagogy and technology of distributed learning. Here members become responsible for and have the tools to “bind together in a social entity” where they share routines, sensibilities, artifacts, vocabulary, and style.

As learners collaboratively build their learning community of practice through the construction of artifacts, vocabulary, routines, sensibilities, and style, they are building shared knowledge.

The idea of building shared knowledge is at the heart of the distributed cognition work of Hutchins (2000). Distributed cognition theory suggests that knowledge and knowledge acquisition go beyond the individual. Rather, knowledge is constructed through the social interactions, where individuals share and co-construct knowledge. In other words, knowledge is not constructed by the individual alone, but instead by the interacting members engaged in the learning process. This is an important characteristic of distributed learning that supports the concept of communities of practice producing a “shared repertoire of communal resources (routines, sensibilities, artifacts, vocabulary, styles, etc.) that members have developed over time” (p. 54). If the joint enterprise of that online group is that of learning, then membership within that community supports and fosters shared knowledge construction. This suggests that, in terms of community building, through their group interactions the social entity and its members build knowledge as they build community.

This section has focused on the elements of community building within online groups, emphasizing that through the development of an online community of practice where group interaction is enhanced by members negotiating the communication norms and behaviors and working collaboratively to build and distribute knowledge. In doing so, through enhanced group interactions the social entity builds community, shares elements and resources that support the individuals within that community. Therefore, to support the initial building of community, it is suggested that administrators and developers of those groups provide ample tools and, perhaps structure, but that it is ultimately the membership that makes the rules. This does not fully address issues of power, as individuals who tend to be more aggressive may still dominate discussions.
and communications. Although the CMC administrators of the online community could censure inappropriate individuals, and some communities may decide that some form of policing is appropriate, still, here, the membership should be responsible for the decision making of that policy. It is recommended that by building an online community of practice, inequities, and disagreements would be settled by the membership of that community, based upon that group’s agreed-upon norms. The next section will outline various communication strategies that will help reduce miscommunications that can often be the cause for online disagreements.

**Utilizing CMC Strategies**

Building a community that shares and co-constructs knowledge could be enhanced by what has been learned about CMC strategies. Although, as has been argued above, the membership of an online community is responsible for setting the communication norms of their group, it still seems prudent to be aware of helpful communication strategies. For example, if the membership of the online community creates routines, sensibilities, artifacts, vocabulary, styles, as Wenger (1998) suggests, it would seem that communicators would benefit from understanding, being aware of, and utilizing CMC strategies within their messages that support more personal interactions. By doing so, there may be less miscommunication, which can often erupt into disagreements.

To better understand communication interactions, Fahy, Crawford, and Alley (2001) developed the transcription analysis tool (TAT), which provides five CMC interaction categories, three of which have subcategories.

- **T1 – Questioning**
  - T1 (a) vertical
  - T1 (b) horizontal

- **T2 – Statements**
  - T2 (a) direct
  - T2 (b) answers or comments

- **T3 – Reflections**

- **T4 – Scaffolding**

- **T5 – References, Authorities**
  - T5 (a) references, quotations, paraphrases
  - T5 (b) citations or attributions

In a later study, Fahy (2003) utilized a specific subset of the categories from the TAT, listed above, to investigate participant use of supportive communication strategies. The supportive communication strategies investigated included:

- **T1 (b) horizontal questions**, which promote more discussion as this type of question does not presume one correct answer, but rather invites any plausible answer.
- **T2 (b) referential statements**, message statements that directly or indirectly refer to another’s message, thus engaging more than just the author of the message.
- **T4 scaffolding and engaging comments** including:
  - **Acknowledgement**: Recognizing or acknowledging comments or statements from prior messages.
  - **Agreement**: Expressing agreement.
  - **Apology, or self-criticism**: Making an apologetic statement.
  - **Closing**: Using some sort of summary or closing (e.g., later).
  - **Emoticon**: Using emoticons (e.g., ;-)).
  - **Humor**: Using humor to lighten the mood.
  - **Invitations**: Statements that invite a response.
  - **Rhetorical questions**: Posing a rhetorical question.
  - **Salutation**: Using a greeting, especially with the individual’s name.
  - **Signature**: Using a signature or nickname.
  - **Thanks**: Expressing thanks.
He found that participants believed that supportive CMC strategies were important and utilized them within an online learning environment in 25.9% of the interactions. In particular, he found that the most commonly used supportive communication strategies utilized by his participants included: (a) referential statements, (b) signatures, (c) greetings, and (d) horizontal questions.

Within a community of practice or learning, it is recommended that members become aware of and begin to utilize these strategies. While it might seem logical to use some of these strategies in group interactions, messages are often received that do not include any closing and, at times, not even something as important as the sender’s name. In utilizing these strategies the CMC message will include more communication cues to help enrich the medium with more information. As described earlier, a lack of nonverbal cues can be a possible issue for female members of a community, and this simple strategy might better support a member’s understanding of a particular message. Simply utilizing emoticons to replace the nonverbal cues (e.g., facial expressions) missing from CMC can provide more information to the recipient of the message. For example, utilizing the winking emoticon can provide information to the recipient that the statement was not meant to be taken seriously. Other more complex strategies, such as stating acknowledgement and agreement, provide the recipient with information that might replace a head nod or other gestures, not communicated in text-only CMC. Furthermore, apologies for miscommunications or other miscues provide “I” statements and self-disclosures that are associated with higher group development communication style (HCS). The use of horizontal questions, inviting responses, and providing referential statements of another’s message support the notion of enhancing social interactions, which encourages additional participant engagement in group interaction. With the use of these more supportive CMC strategies, members will engage in richer CMC to better understand one another, and, thus, augment community building. Again, it is the membership of the community that decides upon the communication norms of that community. However, with more members of the online community being aware of and utilizing these communication strategies the membership is more apt to adopt them, making them a routine of their social entity’s repertoire of vocabulary and style.

Although the various CMC strategies adopted and utilized by a community might improve text-based communication, advances in Internet technologies have brought about the development of multimedia CMC tools also could improve communication. The next section will deal with some of types of current multimedia tools that can better support the needs and predilection of the community.

**Multimedia CMC Tools**

Although computer desktop video conferencing and telephony have increased the media options for CMC, they still do not provide all of the rich aspects of face-to-face interaction. However, these multimedia technologies do add additional communication cues beyond the text-based communication strategies mentioned above. Therefore, it does seem prudent and is recommended to utilize advanced online multimedia communication technologies to the traditional tool set to thicken the communication conduit when possible and, perhaps more importantly, when appropriate.

By utilizing a video conferencing, for example, communication can be enhanced with facial expressions, hand gestures, voice inflection and other cues missing from text-only CMC. In addition, many of these tools provide interactive whiteboards for shared graphic demonstrations. These tools can be used to provide more graphic elements to enhance one’s ideas or concepts. Furthermore, some of these tools also permit document or application sharing. This can be a
valuable tool that can provide technical support and foster online collaboration.

Although these tools can provide the added communication support, they also can complicate the communication process. In terms of video conferencing, for example, some individuals may dislike being on display and this use may cause some discomfort. Also, utilizing the audio and video components, one must have a camera and microphone for each individual, which requires the users, to some extent, to be more technically proficient to set up and use these multimedia peripherals. In addition, high speed Internet connectivity for all parties interacting within the group communication is really a must. Although it is possible to utilize these tools with low bandwidth, it can become very stilted and, thus, very frustrating, perhaps negating the advantages. Furthermore, some of these multimedia CMC tools are open systems allowing one to communicate with anyone, while others are more closed to specific group membership. Because the various types of systems, it is important to know the strengths and weaknesses, and to understand their relative costs and benefits, as they have specific requirements and individual features—keeping in mind the purpose or objective of the group’s communication. The multimedia CMC tools listed below were chosen to provide examples of the different types currently available and are not inclusive of all tools.

**Skype**

Skype™ (n.d.) is an Internet telephony service that utilizes a computer with Internet connectivity, primarily for audio, provides computer to computer Internet Telephony. Although it does provide a video option, it currently is in the beta version. However, once connected users also have a chat window available for synchronous messaging as well. Users can set up conference calls with up to four participants plus the host.

Skype client software includes all of the features one might expect in such a tool, such as an address book of contacts for organizational and ease of use purposes. Skype accounts also provide several services for various fees, such as SkypeIn, which provides the user a phone number much like a tradition phone including voicemail, and Skype-Out, which permits users to make phone calls to both land and cell phone lines, including SMS, a form of instant messaging. Requirements include a headset with microphone, which is used instead of a microphone and speakers to keep feedback from occurring. The easy-to-use client software can be downloaded for free. More importantly, is that computer-to-computer communication is free. There is a fee to utilize calls to and from land or cell phones; users need a pre-paid account. However, the costs are quite inexpensive, especially for international calls. Because this form of Internet Telephony primarily focuses on audio, Skype might be a good tool to consider when beginning utilizing multimedia CMC, as it only requires a headset and the easy-to-use free software to get started.

**iVisit**

iVisit (n.d.) combines a variety of CMC tools including, video conferencing, instant messaging, file sharing, and desktop sharing. Although server licenses are available, most users have to obtain access to the software application in the traditional server/client relationship, where iVisit provides the server access, and users download client software and sign up for one of two types of accounts, Light and Plus. The software includes several windows, including: the video feed with controls, an address book that lists both people and places (available conferences), and a chat window for both the individual and all chatting within a conference.

An iVisit Light is a free account that permits users to interact as guests in open or public con-
ference rooms or invited into private conference rooms. iVisit Plus accounts are yearly subscriptions that permit the ability to host a conference room that can accommodate eight visitors, including the host. Although this can be increased to 16 simultaneous users for no extra fee, it is not recommended due to the complexity of managing that many in an online video conference. In addition to being able to host a conference, the iVisit Plus account provides greater audio quality and video resolution, plus other benefits. One rather interesting feature of iVisit is the ability to make recordings of participants audio and video feeds. Although the Light account only permits saving the file into an iVisit proprietary file format, the Plus account provides for QuickTime recordings. While using iVisit for multimedia, CMC provides greater cues that video provides, it also is more complex to setup, requiring a camera, microphone, and/or headset. Depending upon the technical sophistication and needs of the particular community, this tool may be worth the complexity.

**Microsoft NetMeeting and Messenger**

For quite a few versions, Microsoft Windows-based operating systems have provided multimedia CMC tools that include various features depending upon the particular version. The tools range from earlier versions of Microsoft NetMeeting (n.d.) to, most recently, Microsoft Live Messenger (n.d.). While Live Messenger is quite similar to Yahoo Messenger and focuses on instant messaging (IM), audio-video Internet telephony (somewhat like Skype), and file sharing, NetMeeting also includes application sharing. NetMeeting, will no longer be part of the new Windows Vista operating system. Instead, Microsoft will include a new suite of CMC tools called, Windows Meeting Space for home and home office users, Office Live Meeting Windows Meeting Space for small and medium business users, and Office Live Meeting Office Communications Server 2007 for large enterprises (Microsoft NetMeeting, n.d.). Each level of the application will work similarly to NetMeeting in that it will still support desktop, application, and file sharing across networks. One of the biggest changes from earlier versions of NetMeeting, however, will be the increase in the number of users that can connect simultaneously. One of the biggest advantages of the Microsoft multimedia CMC tool set is that it is readily available within one’s operating system. However, with the newer versions, outlined on the Microsoft Web site, members within online communities will need to decide which tool best suits their needs and available resources.

**Elluminate®**

Elluminate® (n.d.) is a comprehensive online multimedia-conferencing tool that includes audio, chat, interactive whiteboard, application sharing, file transfer, and direct messaging. It truly goes beyond the tools previously listed, as its primary focus is to support group interactions. As such, Elluminate includes additional features designed to provide greater user interactivity, such as: (a) participant profiles to provide greater information (photo & bio) about the participants, (b) polling, which permits a moderator to get some feedback regarding an issue, question, or comment, and (c) breakout rooms, a great feature should users want to break up a large conference for smaller group discussions. One important feature is that it utilizes full duplex audio, meaning that more than one individual can speak at the same time. However, the tool also provides for moderator control, should that be desired. In addition, the tool is designed for assistive access with closed-caption transcript and key stroke configurability. Elluminate also provides for input from a Web cam, which provides video feed of the speaker, although this feature is not the primary focus as with iVisit.
Elluminate provides multimedia CMC on a much more enterprise scale, and, as such, the cost is commensurate. There are several versions available, each appropriately priced, including Lite, Academic, and Enterprise editions, depending upon the user’s size and needs. Currently, Elluminate is being utilized by corporate trainers and universities. In addition, Elluminate has partnered with WebCT Vista, eCollege, Blackboard, and others to devise power links that permit Elluminate to work seamlessly within these course management systems. This multimedia CMC tool is designed for communities that want to focus on the synchronous interactive components that this tool provides. The software also provides for asynchronous delivery by including a session recording feature, which makes sessions available for later viewing. Although Web-delivered video via a Web cam is available, Elluminate does not focus on multiple, simultaneous video feeds such as iVisit. Depending upon need, this tool might provide users with a very interactive multimedia CMC tool, but as with any complex tool, deployment and implementation are a consideration. Again with this tool, users will be required to have a microphone, speakers, or headset to interact with the multimedia elements.

**Multimedia CMC Uses for Support and Collaboration**

The level of multimedia CMC tools available varies from those that are very simple to those that are much more comprehensive and quite very complex. As mentioned, it should be the membership of the community that selects the CMC tools. In addition to increasing the number of communication cues engaged users receive due to the addition of audio and video, there are further ways to utilize the multimedia CMC tools. In particular, the application-sharing feature of some multimedia CMC tools better support collaborations that enhance community building and augment the shared construction of knowledge. For example, one of the primary uses for Microsoft Application Sharing is to provide remote technical assistance. When individuals are having difficulties with a computer application they can be supported by sharing that application with a distant help agent. The user can watch and listen as the help agent provides instruction and demonstrates the application elements that caused confusion. In this way the utilization of a multimedia CMC tool for remote assistance can be used to help members by any more knowledgeable member within the community.

A more collaborative example of utilizing multimedia CMC tool to support a distributed learning environment is using the application sharing for group authorship. In this case, team members can schedule synchronous online meetings where they utilize application sharing to show and discuss Web resources. They also can use application sharing to work collaboratively on the same file to produce a shared Word document or presentation. In doing so, they are collaboratively engaging in the production of an artifact, thus, co-constructing knowledge.

**Summary**

This section provided recommendations that could enhance CMC group interactions. Specifically, it is recommended that the users of CMC consider fostering the building online communities of practice where members co-construct, hopefully, equitable communication norms in their online group interactions. It also is recommended that the members engaged in group interactions become aware of and utilize CMC strategies that enhance their interactions by fostering more clear communications. Finally, it is recommended that multimedia CMC tools be considered to enhance and enrich CMC when practical and appropriate.
FUTURE TRENDS

This chapter has discussed the issues and recommendations of current CMC group interactions. Over the past few years, incredible advances in communication technologies have become available and have been embraced by our society. So what does the future hold? In reviewing the literature of those who’ve studied aspects, issues, and elements of CMC group interaction, one observation is evident—electronic computer-mediated communication has expanded in use, sophistication, and reach. More people now use various types of computer-mediated communication than ever before to communicate with others in all parts of the world. However, the literature provides evidence that online group interactions can suffer the same issues as traditional face-to-face group interactions. With that in mind, one could predict that future trends in CMC group interaction will include the continued expanding use, advancement, and sophistication of CMC tools to an even more global scale than currently exists. With the expanding global reach, users’ skills, in terms of more sophisticated tools and more strategic use, will need to increase as well. With more diverse users making use of the various CMC tools, communicators will need to become better accomplished CMC users to compensate for the limitations of the various CMC systems. In addition, they will need to be more knowledgeable about possible group interaction issues that pertain to newer systems. If that prediction holds true, future global CMC users would benefit from greater investigation of cultural communication differences and global CMC interactions as they relate to the various media that transmits their messages.

Expanding and Changing CMC Skills

Computer-mediated communication has become an important aspect of our society. However, as detailed in this chapter, there are issues that are embedded within CMC group interactions that may impact one’s ability to clearly communicate via the various asynchronous and synchronous systems. It would seem that being aware of the strengths and limitations of the CMC tools and having skill using those strategies, would provide an advantage. For example, having knowledge of, monitoring, and then intentionally utilizing some of the communication strategies listed above may make one more apt to better communicate within a lean or limited CMC system. By utilizing some of the CMC text-based strategies listed above, regardless of gender, one could promote his or her higher group communication style (HCS), thus enhancing the message with richer, more engaging levels of communication. In other words, those individuals who are more CMC savvy, in terms of being able to more purposefully utilize communication strategies would be better able to foster and support persuasive arguments, promote agenda, and clearly share their ideas. Given knowledge of the specific strengths and weaknesses of various communication styles within specific CMC media would provide a powerful advantage to those who might strategically adjust their own communication style accordingly.

Parallel events are beginning to happen within the use of mobile phones for text messaging. For example, users are utilizing text messaging with their mobile phones or instant messaging on their computers for quicker, asynchronous communication when they cannot talk aloud or in lieu of e-mail or other slower CMC systems. When thinking about future possible trends, it seems logical to focus future practical efforts toward educating members of communities engaged in CMC to better understand the limitations of various CMC systems and how best to understand, monitor, and promote clear, concise, and perhaps strategic CMC style. This may become much more important as higher-speed Internet bandwidth becomes more widely available making multimedia CMC tools more common place.
Expanding Need for Global Understandings

This chapter has focused upon detailing specific issues of CMC style differences and media issues, and has provided some recommendations on building communities of practice within these CMC systems. For the most part, the discussion has focused on the research literature that discusses CMC style issues as they relate to group interaction differences and the various media available that might impact the communication cues available. As education, business, and industry extend their reach to more global markets, CMC will be utilized more extensively for interaction by members of various cultures and languages. While certain elements of CMC may better support these endeavors, there will be other issues that could compound good CMC group interaction. For example, asynchronous CMC better supports members in various time zones allowing for business to operate at their normally desired time. With more synchronous multimedia utilization globally, users will be required to interact 24/7. In addition, more global group interactions will require users to be sensitive to diversity in both language and culture. For example, an international student from China was concerned when his advisor sent him an e-mail where some of the text was colored red. He thought she was angry with him—when in truth, she simply was trying to emphasize a point by making it stand out. From her perspective, this was a very benign strategy. Because the color red has specific connotations within the Chinese culture, there was a pretty serious cultural miscommunication for this Chinese student. While no harm was intended by this communication strategy, its use created an unintended consequence of causing great alarm for the student.

Future research in CMC utilization will continue to include the issues of gender and communication style. However, these issues are bound to be expanded by cultural differences as CMC tools are used by members across a more global society trying to build communities of practice with various cultural differences. A very powerful example of building community within a multi-ethnic online discussion group is described in the work of Kadende-Kaiser (2000). She examined CMC interactions of members of Burundinet, a multiethnic discussion forum for Burundians in the Diaspora. Most of the discussion focused on the terrible ethnic violence between the Hutus and Tutsis by members that spoke Kirundi (the native language), French (the colonial language), and English. In addition to their language differences, members came from very different perspectives on a very divisive topic. In this example, social rules were changed to provide more equitable social interactions. “This would not be the case in Burundian social settings, where seniority and social status determine who can interact with whom, or who is even excluded in the interaction. . . . In internet communication, however, these cultural rules regarding turn-taking on the basis of age and social rank are irrelevant; any subscriber can play the role of social mediator” (p. 131). Indeed, the membership of Burundinet fashioned the rules as she states:

"Mutual respect and politeness, on the one hand, and the freedom of expression, on the other, do not have to be mutually exclusive. Therefore, many Burundinet members seek to preserve individual subscriber rights to express their opinion, regardless of what views they have in mind. As one net member advised: ‘It seems more wise that we try to preserve a minimum of formal politeness toward the others, and at the same time we need to allow every Burundinetter to express his/her opinion on Burundi, regardless of what this opinion is’ (Author I, 1996). These statements and individual opinions serve as the foundation for message composition and they contribute to the establishment of the rules for interaction on Burundinet. (p. 134)"
Although cultural differences, such as this, may cause issues when filtered through the lean medium of current CMC technologies within a more global community, the membership of Burundinet demonstrated how it established unique rules for CMC interaction within their virtual community to permit discussion regardless of their differences. This provides a very good example of how the CMC group interaction communication norms are established by the membership.

Summary

This section provided some ideas of what the future might hold for CMC group interactions. Certainly, we can presume that CMC usage will increase and become more complex and, perhaps, robust. CMC users will be required to become more technically proficient. A more important concern, however, will be that as the usage of CMC expands more globally, the communities that will better prosper will be those that trend toward the development of communities of practice, where members think about their responsibility in the development of their shared repertoire of communal resources in light of possible cultural differences. It will be vital for members engaged in CMC group interactions to be aware of, and provide consideration for, cultural differences. While the CMC strategies outlined above may work well for certain cultures, will they work well within a global community? Or, will they have unintended consequences?

Future Research Directions

There is a history of research that has investigated face-to-face group interaction that supersedes the CMC tools that are now employed. The findings from that earlier research seems to parallel research findings of those engaged in CMC group interactions. For example, gender-based CMC style differences can impact those engaged in online group interactions. However, there are strategies that users can and do employ to adjust their CMC group interactions in an attempt to replace missing nonverbal cues, thus supporting and fostering better communication. Therefore, future research that further investigates how users manipulate their CMC style strategically by employing CMC techniques and/or adjusting their CMC group interactions, based upon the medium, could prove very enlightening. Because members engaging in CMC group interactions can and do take responsibility for developing their community of practice, researchers also would be advised to study the process of defining and developing communication norms within an online community of practice.

As technologies continue to advance, richer media will become commonplace for CMC group interactions. Regardless of new, more advanced multimedia electronic communication systems, there still will be a filtering of sorts on communication cues within CMC systems. Future researchers would be advised to focus their investigations on how multimedia might hinder or enhance group interactions based upon gender, and perhaps more importantly, culture. In particular, it seems essential to investigate global CMC group interactions in light of cultural and language differences. For example, community systems like BuddySpace, that provide the users’ location could provide cues to support those engaging in group interactions within that type of online community space.

Researchers should be encouraged to take advantage of the known CMC issues and focus their investigations on more global issues that are bound to emerge as communities expand. While gaining knowledge of how more robust, multimedia CMC might impact various individuals it vital, future research should go beyond and also focus on how to best educate members of global communities of practice to provide clear, concise communication within those more advanced electronic communication systems. With continued knowledge of the advanced CMC systems and the diverse
cultural interactions that happen with them, this form of information technology will better serve the expanding global community.

Finally, instant messaging (IM) and text messaging has become widespread, globally. This form of asynchronous communication is novel in that it changes the type and, perhaps, the reason for a particular communiqué. For example, one might “text” or “IM” another because they are busy—working, in class, or talking with others—and cannot be interrupted with a traditional phone call or chat. Furthermore, this form of communication has developed its own lexicon and syntax. While people may have wanted to communicate in similar ways in the past, this technology has changed the way many communicate, as evidenced by its overwhelming use. Specifically, this technology has provided the opportunity for people to multitask in many areas of life, which brings up an interesting question. How will this new form of communication impact or change more traditional forms of communication? With that in mind, an interesting opportunity for future research would be to investigate how technology advances have impacted or changed more historical forms of human communication and social interactions.

REFERENCES


Fahy, P. J. (2003). Indicators of support in online interaction. *International Review of Research in Open and Distance Learning, 4*(1).


Computer-Mediated Communication


**ADDITIONAL READING**


Chapter 4.9
Mobile Phone Use Across Cultures: A Comparison Between the United Kingdom and Sudan

Ishraga Khattab
Brunel University, UK

Steve Love
Brunel University, UK

ABSTRACT

Recently, the ubiquitous use of mobile phones by people from different cultures has grown enormously. For example, mobile phones are used to perform both private and business conversations. In many cases, mobile phone conversations take place in public places. In this article, we attempt to understand if cultural differences influence the way people use their mobile phones in public places. The material considered here draws on the existing literature of mobile phones, and quantitative and qualitative work carried out in the UK (as a mature mobile phone market) and the Sudan (that is part of Africa and the Middle East culture with its emerging mobile phone market). Results indicate that people in the Sudan are less likely to use their mobile phones on public transport or whilst walking down the street, in comparison to their UK counterparts. In addition, the Sudanese are more willing to switch off their mobile phones in places of worship, classes, and meetings. Implications are drawn from the study for the design of mobile phones for different cultures.

INTRODUCTION

Economic globalization and the widespread use of mobile phones have changed the way people live and manage their lives, and cut down the virtual distance between countries, regions, and time zones. New ways of using mobile phones are constantly emerging (e.g., downloading music to listen to on the train), and the pervasive use of mobile phones in public places for private talk...
Mobile Phone Use Across Cultures

(both business- and socially-oriented) is a clear example of how mobile phones are changing our economic and social lives. As a result of this, there is an emergent body of research on the use of mobile phones in social spaces. For example, Ling (2004) highlights how their use in public places has raised questions of what the appropriate or inappropriate behaviour is in public places. In this study, he found that people perceived mobile phone use in places such as restaurants as unacceptable, partly because mobile phone users tend to talk louder than usual so that people nearby feel intruded upon, embarrassed, and have a sense of being coerced into the role of eavesdropper on a private conversation.

Research has also shown that mobile phones can occupy concurrent social spaces, spaces with behavioural norms that sometimes conflict, such as the space of the mobile phone user, and the virtual space where the conversation takes place (Palen, Salzman, & Youngs, 2000). This feeling of conflict has led researchers in this area to propose that the use of mobile technology in public places is creating a new mixture of public and private space that has yet to be accommodated by for users of mobile technology and bystanders in terms of what is acceptable or unacceptable behaviour.

This phenomenon has been analysed predominately using concepts drawn from Goffman’s analysis of social interaction in public places (Goffman, 1963). In this work, Goffman suggested that people have specific “public faces” and personas for different public social locations. The idea behind this is that individuals have rules that determine their behaviour in public places, or what Burns (1992) refers to as the “observance of social propriety.” For example, Murtagh (2001) presented findings from an observational study of the nonverbal aspects of mobile phone use in a train carriage. Murtagh found that changing the direction of one’s gaze—turning one’s head and upper body away from the other people sitting next to you in the carriage—was a common feature of mobile phone behaviour on trains. These behavioural responses were seen as being indicative of the subtle complexities involved when using mobile phones in public locations. This study suggests that mobile phone users are actively engaged in trying to distance themselves from their current physical location in order to enter a virtual environment with the person they are having a mobile phone conversation. In relation to this, Love and Perry (2004) used role-play experiments to investigate the behaviour and attitudes of bystanders to a mobile phone conversation. They found that participants had strong views on embarrassment, discomfort, and rudeness. They also report that the actions of those who were put in the position of overhearers followed a pattern: they acted as though they were demonstrably not attending, even though they were all able to report accurately on the content of the conversation.

However, to date, most of the research reported in this area has tended to focus on what is termed the developed world. Mobile phones are also transforming people’s lives in the developing world. In Africa, the unreliable and inefficient landline telecommunication infrastructure has made the mobile phone the solitary available communication tool for many people (BBC, 2003). However, as mobile phone use in Africa continues to grow, there is a need for mobile phone companies who are entering this market to consider the possible impact of cross-cultural differences in people’s attitude towards mobile phone and service applications.

This article first briefly reviews relevant literature about the use of mobile phones in public places. The concept of culture and cultural models are explored in the second section. In the third section, the methods of this study are presented. Techniques of collecting the data and the procedure of this study are presented in the fourth and fifth sections, respectively. Some key findings from the study are presented and discussed in the
sixth and seventh sections with reference to how cultural differences might affect mobile phone use in public places. Finally, the conclusion of this study is presented in the last section.

WHAT IS CULTURE?

Culture is a complicated paradigm that is difficult to accurately define. According to some researchers, culture must be interpreted (van Peursson, in Evers & Day, 1997). Hofstede (1980) conceptualized culture as “programming of the mind,” suggesting that certain reactions were more likely in certain cultures than in others, based on differences between the basic values of the members of different cultures (Smith, Dunckley, French, Minocha, & Chang, 2004). Culture can also be seen as a collection of attributes people acquire from their childhood training. These attributes are associated with their environment, surroundings that influence the responses of people in that culture to new ideas, and practices and use of new technology (such as mobile phones). Given that culture may affect the way people behave and interact in general, Ciborowski (1979) identified a close link between knowledge and culture. In the context of mobile phone communication, it may be argued that culture influences knowledge—or the individual’s general experience—therefore affecting, in this instance, their attitude towards mobile phone use in public places.

Another explanation of culture has been offered by Hofstede (1980). He produced a cultural model that focuses on determining the patterns of thinking, feeling, and acting that form a culture’s “mental programming.” This model has been adopted for the study reported in this article, as researchers in the area of cross-cultural differences and technology use consider it a valid and useful measure of systematic categorization (e.g., De Mooij, 2003; Honald, 1999). In addition, it is also considered to be directly related to the relationship between product design and user behaviour (De Mooij & Hofstede, 2002). An explanation of Hofstede’s cultural dimensions is as follows:

- **Power distance**: the extent to which less powerful members expect and agree to unequal power distribution within a culture. The two aspects of this dimension are high and low power distance.

- **Uncertainty avoidance**: discusses the way people cope with uncertainty and risk. The two faces of this dimension are high uncertainty avoidance and low uncertainty avoidance.

- **Masculinity vs. femininity**: refers to gender roles, in contrast to physical characteristics, and is usually regarded by the levels of assertiveness or tenderness in the user. The two aspects of this dimension are masculinity and femininity.

- **Individualism vs. collectivism**: deals with the role of the individual and the group, and is defined by the level of ties between an individual in a society. The two aspects of this dimension are individualism and collectivism.

- **Time orientation**: deals with the extent to which people relate to the past, present, and future. The two aspects of this dimension are short-term orientation and long-term orientation.

A number of cross-cultural studies have investigated differences in attitudes towards new technology. Smith, French, Chang, and McNeill (2001) carried out a study using Hofstede’s model. They adapted the Taguchi method—a partial factorial experimental design method—in order to investigate differences between British and Chinese users’ satisfaction and preferences for Web sites. They found significant differences between British and Chinese users in their prefer-
ence of detailed e-finance product information. For example, Chinese users tended to adopt a more holistic approach to viewing Web content, as compared to British users.

In another study, Honald (1999) found that German mobile phone users preferred clearly-written and inclusively rich user manuals, whereas Chinese mobile phone users focused on the quality of the pictorial information.

Evers and Day (1997) found that there are clear cultural differences between user acceptance of interfaces for different cultural groups. In their study, they found differences between Chinese and Indonesian users. Indonesians were found to like soft colours, black and white displays, and pop-up menus more than Chinese users. Also, Indonesians seemed to prefer alternative input and output modes (e.g., sounds, touch screens, data gloves, and multimedia) in comparison to the Chinese who preferred the use of many different colours for the interface design.

Despite the importance and the relevance of cultural factors and its impact on the use of global products and services (such as mobile phones), little research has compared the effect of cultural differences on issues such as social usability of mobile phone use in the developing and the developed world. Sun (2003) argues that variation in cultural states will cause different attitudes or ways of using mobile phones.

The practice of the “missed call” is a clear example of how users from different cultures develop their own usage style. The missed call is when the caller places a mobile phone call and purposely hangs up before the recipient can answer the call. Donner (2005) investigated the phenomenon in Sub-Saharan Africa where the missed call is known as “Beeping.” He found that users have produced elaborated codes and social messages to be exchanged over the network without bearing any cost—or at least not from those who are in a less secure financial situation.

Another exclusive mobile phone cultural attitude is evident in Bangladesh, Uganda, and Rwanda, where a woman, for example, borrows money to buy a special mobile phone designed for multiple user accounts and rural access. After this, she then buys minutes in bulk and resells them to customers in her village. This programme is funded by the Grameen Bank mainly for use in rural areas (Bayes, von Braun, & Akhter, 1999).

The same idea has mutated slightly in the big cities of Egypt and the Sudan where vendors who own fixed contract mobile phones buy a bulk of talk minutes and resell them in smaller chunks to individuals in order to make a profit. This practice is accommodated by their national phone service providers and is known as “balance transfer.” Obviously, this practice cannot be seen in London or any of the developed world cities.

If mobile phone users have different usage patterns, the question that the study in this article addresses is: can we assume that people from different countries use mobile phones in the same way? Thus the question arises: are there any roles for cultural differences in the way people use their mobile phones in public places? Therefore, the attitude of the British (a mature mobile phone user market) and the Sudanese (an emerging mobile phone user market) were examined in relation to their attitudes towards the use of mobile phones in public places.

**METHODOLOGY**

**Participants**

88 participants took part in the study: 43 British (22 male, 21 female) and 45 Sudanese (20 male, 25 female), ranging in age from 15 to 63 years old, with the average age of 30 years. All participants were mobile phone users. The range of mobile phone use for the Sudanese participants was from 2-5 years, whereas the British participants had used mobile phones for 4-12 years.
Data Collection

Data was collected in this study using a questionnaire and an interview. The development of the questionnaire went through several stages. First, the generation of the questionnaire was collated by employing an exhaustive review of the literature generally on mobile phones, human-computer interaction (HCI), and cultural issues in mobile phone use. Second, an in-depth session was conducted with participants from both countries (the UK and the Sudan) to develop the questionnaire. Initially, a total of nine Likert-type questions were developed. The scale was then tested for content validity, which can be defined as the extent to which a test actually measures what it is supposed to measure (Rust & Golombok, 1989). This was undertaken using what is known as the judgemental approach, with three mobile HCI experts.

As a result of this process, the questionnaire was subsequently revised to consist of six Likert-type questions. The six Likert statements focused on attitudes towards the use of mobile phones in public places. An example of the Likert statement used in this study is as follows:

Mobile phones should not be switched off during meetings:

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

The attitude scale had a combination of positive and negative statements in order to control for any possible acquiescence effect from participants when they were completing the attitude questionnaire. This is a phenomenon whereby participants in a study may unwittingly try to respond positively to every question in order to help the investigator with their study. This type of questionnaire format is one of the most common methods used to elicit attitudes from users in HCI research (Love, 2005).

In addition to the questionnaire, a semistructured interview was carried out. The interview questions included open-ended and closed questions, and were designed to gather information on the use of mobile phones in public places, the practice of the missed call, and other features such as the use of mobile phone caller ID. The main points that were covered in the interview were:

1. Attitude towards the use of mobile phones in public places.
2. The use of the missed calls types in the two cultures. For example, the type of missed calls used and the social messages sent through the missed call, and how recipients differentiate between these types of missed calls.

Examples of questions covered in the interview were:

*How do you feel about using mobile phones on public transport?*
*How do you feel about using mobile phones in school during classes?*
*How do you feel about using mobile phones in restaurants?*

**PROCEDURE**

Participants were chosen from an opportunistic sample in both the UK and Sudan and asked to complete the questionnaire and return them to the researcher once they had completed them.

The questionnaires took approximately 15 minutes to complete. At this point, an arrangement was made to interview a subset of the participants who had been selected randomly and volunteered to answer the interview questions. Participants were informed from the outset that the results of
the study would be anonymous, and they would be able to obtain the results of the study from the researcher on request.

RESULTS

An independent sample T test was carried out to compare attitudes towards using mobile phones in public places in the UK and the Sudan. There was a significant difference found in the attitudes for using mobile phones in public transport between the British and the Sudanese ($t=5.99$, $p<0.001$). The British were more willing to use it on public transport than the Sudanese.

Another significant difference was noted between the two countries towards using mobile phones whilst walking on the street. Again, the

| Table 1. Attitudes towards the use of mobile phones in public places in the UK and the Sudan |

<table>
<thead>
<tr>
<th></th>
<th>COUNTRY</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>t</th>
<th>df</th>
<th>P Value Sig 2 tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>I would be comfortable using my mobile phone in restaurants</td>
<td>British</td>
<td>42</td>
<td>2.83</td>
<td>1.146</td>
<td>.177</td>
<td>1.325</td>
<td>70.241</td>
<td>.189</td>
</tr>
<tr>
<td></td>
<td>Sudan</td>
<td>45</td>
<td>2.56</td>
<td>.755</td>
<td>.113</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I would not be comfortable using my mobile phone on public transport</td>
<td>British</td>
<td>42</td>
<td>3.29</td>
<td>1.175</td>
<td>.181</td>
<td>5.925</td>
<td>69.046</td>
<td>.000 ***</td>
</tr>
<tr>
<td></td>
<td>Sudan</td>
<td>45</td>
<td>2.02</td>
<td>.753</td>
<td>.112</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I would be comfortable using my mobile phone whilst walking down the street</td>
<td>British</td>
<td>42</td>
<td>3.69</td>
<td>1.070</td>
<td>.165</td>
<td>3.884</td>
<td>82.171</td>
<td>.000 ***</td>
</tr>
<tr>
<td></td>
<td>Sudan</td>
<td>45</td>
<td>2.84</td>
<td>.952</td>
<td>.142</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobile phones should be switched off in places of worship</td>
<td>British</td>
<td>42</td>
<td>4.45</td>
<td>.861</td>
<td>.133</td>
<td>3.094</td>
<td>51.314</td>
<td>.003 **</td>
</tr>
<tr>
<td></td>
<td>Sudan</td>
<td>45</td>
<td>4.89</td>
<td>.318</td>
<td>.047</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobile phones should not be switched off during meetings</td>
<td>British</td>
<td>42</td>
<td>3.88</td>
<td>.968</td>
<td>.149</td>
<td>2.316</td>
<td>69.411</td>
<td>.023 *</td>
</tr>
<tr>
<td></td>
<td>Sudan</td>
<td>45</td>
<td>4.29</td>
<td>.626</td>
<td>.093</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobile phones should be switched off in schools during classes</td>
<td>British</td>
<td>42</td>
<td>4.00</td>
<td>1.307</td>
<td>.202</td>
<td>2.552</td>
<td>61.278</td>
<td>.013 *</td>
</tr>
<tr>
<td></td>
<td>Sudan</td>
<td>45</td>
<td>4.58</td>
<td>.690</td>
<td>.103</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Mobile Phone Use Across Cultures

British were more favourable towards this than the Sudanese \((t=3.884, p<0.001)\). The Sudanese were found to be more willing to switch off their mobile phones in places of worships, meetings, and in schools during classes. Please see Table 1 for a summary of the main results.

In terms of differences between the attitude of the British and the Sudanese males, an unrelated

Table 2. Attitude difference between the Sudanese males in using mobile phones in public places and the British males

<table>
<thead>
<tr>
<th>Mobile phones should be switched off in places of worship</th>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>t</th>
<th>df</th>
<th>P value sig 2 tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sudanese Male</td>
<td>20</td>
<td>4.90</td>
<td>.308</td>
<td>.069</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>British Male</td>
<td>23</td>
<td>4.43</td>
<td>.992</td>
<td>.207</td>
<td>2.134</td>
<td>26.761</td>
<td>.042</td>
</tr>
<tr>
<td>Mobile phones should be switched off during meetings</td>
<td>Sudanese Male</td>
<td>20</td>
<td>4.20</td>
<td>.523</td>
<td>.117</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>British Male</td>
<td>23</td>
<td>3.83</td>
<td>1.154</td>
<td>.241</td>
<td>1.397</td>
<td>31.583</td>
<td>.172 ***</td>
</tr>
<tr>
<td>Mobile phones not to be switched on in schools during classes</td>
<td>Sudanese Male</td>
<td>20</td>
<td>4.50</td>
<td>.827</td>
<td>.185</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>British Male</td>
<td>23</td>
<td>4.17</td>
<td>1.403</td>
<td>.293</td>
<td>.942</td>
<td>36.374</td>
<td>.352</td>
</tr>
<tr>
<td>I would be happy using mobile phones in restaurants</td>
<td>Sudanese Male</td>
<td>20</td>
<td>2.50</td>
<td>.688</td>
<td>.154</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>British Male</td>
<td>23</td>
<td>2.70</td>
<td>1.105</td>
<td>.230</td>
<td>-.706</td>
<td>37.389</td>
<td>.485 *</td>
</tr>
<tr>
<td>I would not be comfortable using a mobile phone on public transport</td>
<td>Sudanese Male</td>
<td>20</td>
<td>2.25</td>
<td>.786</td>
<td>.176</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>British Male</td>
<td>23</td>
<td>3.13</td>
<td>1.180</td>
<td>.246</td>
<td>-2.912</td>
<td>38.570</td>
<td>.006 **</td>
</tr>
<tr>
<td>I would be comfortable using a mobile phone whilst walking on the street</td>
<td>Sudanese Male</td>
<td>20</td>
<td>3.15</td>
<td>.813</td>
<td>.182</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>British Male</td>
<td>23</td>
<td>4.04</td>
<td>.825</td>
<td>.172</td>
<td>.869</td>
<td>40.330</td>
<td>.001 **</td>
</tr>
</tbody>
</table>

*\(P<0.05\); **\(P<0.01\); ***\(P<0.001\)
T test revealed that the British males are more willing to use mobile phones on public transport and when walking on the street than the Sudanese males ($t=-2.912$, $t=.869$, $p<.001$). Please see Table 2 for a full summary of the results.

Comparing the attitudes of the British and the Sudanese females towards the use of mobile phones in public places—an unrelated T test revealed the British females are more relaxed using mobile phones in public transport than the

<table>
<thead>
<tr>
<th>Mobile phones should be switched off in places of worship</th>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>$t$</th>
<th>df</th>
<th>P value sig 2 tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUDANESE FEMALE</td>
<td>25</td>
<td>2.60</td>
<td>.816</td>
<td>.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRITISH FEMALE</td>
<td>19</td>
<td>3.00</td>
<td>1.202</td>
<td>.276</td>
<td>-1.248</td>
<td>30.068</td>
<td>.222</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mobile phones should be switched off during meetings</th>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>$t$</th>
<th>df</th>
<th>P value sig 2 tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUDANESE FEMALE</td>
<td>25</td>
<td>1.84</td>
<td>.688</td>
<td>.138</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRITISH FEMALE</td>
<td>19</td>
<td>3.47</td>
<td>1.172</td>
<td>.269</td>
<td>-5.408</td>
<td>27.256</td>
<td>.000 ***</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mobile phones should not be switched in schools during classes</th>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>$t$</th>
<th>df</th>
<th>P value sig 2 tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUDANESE FEMALE</td>
<td>25</td>
<td>2.60</td>
<td>1.000</td>
<td>.200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRITISH FEMALE</td>
<td>19</td>
<td>3.26</td>
<td>1.195</td>
<td>.274</td>
<td>-1.955</td>
<td>34.863</td>
<td>.059</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I would be happy to use my mobile phone in a restaurant</th>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>$t$</th>
<th>df</th>
<th>P value sig 2 tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUDANESE FEMALE</td>
<td>25</td>
<td>4.88</td>
<td>.332</td>
<td>.066</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRITISH FEMALE</td>
<td>19</td>
<td>4.47</td>
<td>.697</td>
<td>.160</td>
<td>2.348</td>
<td>24.196</td>
<td>.027 *</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I would not be comfortable using a mobile phone on public transport</th>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>$t$</th>
<th>df</th>
<th>P value sig 2 tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUDANESE FEMALE</td>
<td>25</td>
<td>4.36</td>
<td>.700</td>
<td>.140</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRITISH FEMALE</td>
<td>19</td>
<td>3.95</td>
<td>.705</td>
<td>.162</td>
<td>1.929</td>
<td>38.758</td>
<td>.061</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I would be comfortable using a mobile phone whilst walking on the street</th>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>$t$</th>
<th>df</th>
<th>P value sig 2 tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUDANESE FEMALE</td>
<td>25</td>
<td>4.64</td>
<td>.569</td>
<td>.114</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRITISH FEMALE</td>
<td>19</td>
<td>3.79</td>
<td>1.182</td>
<td>.271</td>
<td>2.892</td>
<td>24.322</td>
<td>.008 **</td>
<td></td>
</tr>
</tbody>
</table>

*P<0.05; **P<0.01; ***P<0.001
Sudanese females ($t=2.348, p<.001$). Please see Table 3 for a full summary of the results.

**INTERVIEW RESULTS**

The interview results corresponded with the questionnaire data, indicating that there is a difference between the British and the Sudanese attitudes towards the use of mobile phones in public places. Sudanese were found to be less willing to use mobile phones in public places than their British counterparts. In the interview, Sudanese participants revealed various reasons for their uncomfortable attitude towards the use of mobile phones in public places. For example, some of the participants felt that the use of mobile phones in public transport is unacceptable because it can be disturbing to other people in close proximity to the mobile phone user. As one of the Sudanese interviewees commented:

*Using a mobile phone in public places, especially on public transport where you are closely surrounded by people, is not something that you can do comfortably. It is viewed as improper and unacceptable, as it disturbs others.*

Another Sudanese interviewee added:

*The use of mobile phones on public transport may be considered as a sign of disrespect to others. In particular, to older passengers who you have to respect and act quietly around them.*

An added justification that was revealed by Sudanese participants for not feeling comfortable using mobile phones in public places was related to their tight rules in keeping private issues private, as one of the interviewees commented:

*The use of mobile phones in public places to discuss private matters can put you in an awkward situation; because most of the people surrounding you will hear your conversation and this attitude in itself is not acceptable in our community. People are not expected to discuss private issues so publicly.*

On the other hand, British participants were found to be more comfortable using mobile phones in public places as one of the interviewees commented:

*I have no problems using my mobile phone in public places and especially on public transport, as I can make use of time while sitting there doing nothing.*

Another British interviewee added:

*I use my mobile phone in public places all the time and it does not bother me at all that people are listening to my mobile phone conversations. I do not know them and it is unlikely they are going to know more details about the topic I am discussing.*

The results of this study also indicated that Sudanese females were less willing to use mobile phones in public places than British females. Sudanese females felt that the use of mobile phones in public places, especially on public transport, could attract unwanted attention to them in a society that expects females to keep a low profile. This was echoed in one of the Sudanese female interviewee’s comments:

*I do not like using my mobile phone in public places at all as it only magnetizes others’ attention towards me. If you are on the mobile phone in a public place, people start gazing at you unappreciatively.*

Another Sudanese female interviewee added:
Usually, I do not use my mobile phone in public places. I prefer to keep a low profile. For me, this attitude is a sign of respect for my self and others.

British females appeared to have different view—most of the interviewees were found to feel more comfortable using their mobile phones in public places. As one of the British interviewees commented:

*I prefer to use my mobile phone in public places; it keeps me busy and in a way safe, for example when I want to get my car from the car park when it is dark, I always make sure that I am talking to one of my friends on the mobile phone just in case something happens.*

**DISCUSSION**

The results from the study were interpreted in the light of Hofstede’s cultural dimensions to try and gain some insight into the way culture may influence the use of mobile phones in public places.

It appears from the results that the British generally are more comfortable using mobile phones in public places than their Sudanese participants, who are more reluctant to use mobile phones in contexts such as public transport and whilst walking along the street.

The collectivistic culture to which the Sudan belongs to (using Hofstede’s criteria) indicates an inclination toward a tightly-knit social framework (Hofstede, 1980). The priority is for the groups’ needs, rather than the individual wishes. Therefore, perhaps the use of mobile phones in public places for private talks can be seen as a self-centred act, and quite impertinent for the group needs. The group expects the individual to be considerate to the established social etiquette. The mobile phone user in public transport is expected to adhere to the social protocol and to respect other people’s privacy.

Another reason for the British comfortable attitude to mobile phone use in public places may be due to bystanders’ nonverbal communication attitude. This concept is highlighted by Goffman (1963) where he refers to it as “civil inattention.” Civil inattention refers to the ways in which people acknowledge the existence of others without paying them extra attention; he regarded this as a gesture of respect required from strangers. Lasen (2002a) found that “civil inattention” is clearly present in UK culture: the British tend to avoid open and straightforward looking at other people, and keep away from paying direct attention to others, especially on public transport, such as the Underground. He suggested that this attitude may encourage British mobile phone users to talk more freely outdoors without being concerned about others watching them.

In contrast, in the Sudan, it was noted that “civil inattention” is not clearly evident. Sudanese people tend to look at each other directly. Lasen (2002a) suggested that a lack of proper gaze in certain cultures where “civil inattention” does not rule may be viewed as a lack of respect or ignorance. This lack of civil inattention perhaps justifies the reason behind the Sudanese unwillingness to use their mobile phones in public places, as they are influenced by bystanders’ nonverbal communication attitude. One can say the more civil inattention paid to others, the more free and relaxed they might feel towards using their mobile phones, and vice versa.

Another justification for not using mobile phones in public places might be due to the high score that the Sudan attained on Hofstede’s uncertainly avoidance dimension. According to Hofstede, cultures with high uncertainty avoidance scores tend to be expressive—people talk with their hands, raise their voices, and show emotions. These characteristics can play a role in decreasing the need to carry out private conversations in public places because people in these cultures know that they tend to talk loudly and expressively, which attract bystanders’ attention,
Mobile Phone Use Across Cultures

plus there is a high risk of being known to people around you. Another important point is that as Sudanese people in general talk loudly and in an expressive way, this tends to increase the level of external noise for mobile phone users. Therefore, people talking on mobile phones need to raise their voices more to win over competitive speakers. This loud talking may attract bystanders’ attention and invite eavesdroppers, which can cause a feeling of embarrassment on the part of the mobile phone user. In addition, mobile phone users may feel that bystanders might disrespect them if they discuss their private matters publicly.

Additionally, the Sudanese attitude might be related to the high score obtained on Hofstede’s power distance dimension, where a tight set of social rules are established, and people are expected to follow and respect these rules. For example, the social protocol for behaviour in public places is well recognized in the Sudan, and people are expected to behave in certain ways and not to speak loudly in front of others (especially older people). Private issues should be kept private and dealt with in a private manner and in private settings. It is considered improper to breach these norms. Although in the UK, a social protocol for behaviour in public places also exists, the maturity of the UK mobile phone market may have relaxed or altered people’s expectations and acceptance behaviour in public places. Palen et al. (2000) found that a person’s attitude towards public mobile phone use changes (becomes more accepting) as their mobile use increases. In addition, Palen (2002) predicted that as adoption of mobile phones increases, people will be less disturbed about proper use, but will still prefer to have “mobile free” zones.

In terms of specific gender differences, Sudanese females were found to be more uncomfortable about using mobile phones in public places in comparison to British females. This attitude fits in with the “feminine” attribute of the Sudan culture suggested by Hofstede (1980), where the prevailing value is caring for others. The UK, in contrast, is judged by Hofstede to be more masculine-oriented, and the dominant values are achievement and success.

Although the Sudanese females practice all their rights in terms of education, work, leisure, and the like, they are looked after and cared for by the whole society. As a result of this caring perception towards females in the Sudanese culture, their attitudes and behaviours are more controlled and guarded as they are expected to follow social protocols more than men. For example, Sudanese females are expected to keep a low profile and deflect attention from themselves by talking quietly—and preferably avoid talking—in public spaces.

On the other hand, according to the results of this study, British females are more comfortable using mobile phones in public places. This may be due to the feminine attribute of the UK suggested by Hofstede (1980) where women are seen as equal to men, and they are expected to look after and guard themselves more autonomously. In contrast to the Sudanese females, British females can use mobile phones in public places as “symbolic bodyguards” (Lasen, 2002b). In this context, mobile phones are used as a technique to defend your private space within areas that are heavily populated with unknown strangers (Cooper, 2000; Haddon, 2000). As Goffman (1963) has remarked, women especially do not like to show themselves alone in public places, because this may indicate that they are not in a relationship: a condition which (1) provides a bad impression of their social status and (2) leaves them in a vulnerable situation which can be acted upon by unknown males. To deal with these situations, the mobile phone is quite useful, as it works as a safety net and indicates that this person has their social networks and is not isolated (Plant, 2002).

The other significant result reported in this study is that the Sudanese are more likely to switch off their mobile phones in places of worship. Measuring these results against the Hofstede
typology, the Sudanese score high on uncertainty avoidance scale—religion is valued and greatly respected. People’s attitude towards switching off mobile phones in places of worship in the Sudan is therefore expected. It is also related to the high scores Sudan has on power distance, as roles are set, and religious men are very much valued and respected in the society, so both the Muslim and the Christians in the Sudan tend to be aware of the importance of switching off their mobile phones in places of worship. This result could also be related to the reduced number of people in the UK attending places of worship.

The Sudanese also appear more willing to switch off their mobile phones during meetings than the UK participants. This attitude may be related to their high score in the power distance dimension where people are expected to respect the structure, rules, and the norms of the setting where they are currently present.

As for the British disinclination to switch off their mobile phones during meetings, it might be related to the individualistic feature of the British society, where time is valued, and there is a push for making good use of it. It may also be related to the maturity of British mobile phone adoption where mobile phones have blurred the borders between business and social rules. In relation to this, Churchill (2001) found that the mobile phones in the UK are used to form and maintain both work and leisure relationships.

CONCLUSION

The increased use of mobile phones by people from different cultural backgrounds has become an integral part of our world phenomena, yet to date the impact of cultural differences on the way people use their mobile phones—and its implications on mobile phone design—has failed to be investigated comprehensively. As this article illustrates, mobile phone users with cultural differences were found to use their mobile phones in different ways, and their attitudes may have been influenced by their cultural norms. Although one can argue that cultural norms can be reshaped by technology, results obtained from this study indicate that cultural heritage would appear to influence users’ mobile phone behaviour. The results obtained from this study also suggest that mobile phone designers need to develop a richer understanding of culture in order to develop mobile phones that satisfy culturally specific needs, and thus support mobile phone users’ in their current and potential future communication activities. This is an issue we intend to explore in the next phase of our research.

REFERENCES


Mobile Phone Use Across Cultures

at the Wireless World Workshop, University of Surrey.


Ling, R. (2004). The mobile connection the cell phone’s impact on society. San Francisco: Morgan Kaufmann.


Chapter 4.10
Gender and Computing at University in the UK

Ruth Woodfield
University of Sussex, UK

INTRODUCTION

In the late 1970s, women’s progress and participation in the more traditional scientific and technical fields, such as physics and engineering, was slow, prompting many feminist commentators to conclude that these areas had developed a near-unshakeable masculine bias. Although clearly rooted in the domains of science and technology, the advent of the computer was initially seen to challenge this perspective. It was a novel kind of artefact, a machine that was the subject of its own newly created field: “computer science” (Poster, 1990, p. 147). The fact that it was not quite subsumed within either of its parent realms led commentators to argue that computer science was also somewhat ambiguously positioned in relation to their identity as masculine. As such, it was claimed that its future trajectory as equally masculine could not be assumed, and the field of computing might offer fewer obstacles and more opportunities for women than they had experienced before. Early predictions of how women’s role in relation to information technology would develop were consequently often highly optimistic in tone. Computing was hailed as “sex-blind and colour-blind” (Williams, Cited in Griffiths 1988, p. 145; see also Zientara, 1987) in support of a belief that women would freely enter the educational field, and subsequently the profession, as the 1980s advanced.

During this decade, however, it became increasingly difficult to deny that this optimism was misplaced. The numbers of females undertaking undergraduate courses in the computer sciences stabilised at just over a fifth of each cohort through the 1980s and 1990s, and they were less likely to take them in the more prestigious or research-based universities (Woodfield, 2000).

Tracy Camp’s landmark article “The Incredible Shrinking Pipeline” (1997), using data up to 1994, plotted the fall-off of women in computer science between one educational level and the next in the US. It noted that “a critical point” was the drop-off before bachelor-level study—critical because the loss of women was dramatic, but also because a degree in computer science is often seen as one of the best preparatory qualifications for working
within a professional IT role. The main aim of this article is to examine how the situation has developed since 1994, and within the UK context. It will also consider its potential underlying causes, and possible routes to improvement.

BACKGROUND

In the UK, throughout the 1990s and into the new millennium, the achievements of secondary school-age girls (11-16 years) progressed significantly in the more traditional scientific and technical subjects, and began surpassing those of boys. Before an age when some curriculum choice is permitted (14 years old), girls perform better in science. Furthermore, although fewer of them take science once they have choice, they continue to surpass boys’ achievements in the area. Higher proportions of girls now gain an A-C grade pass in their GCSE examinations in chemistry and biology and physics (Department of Trade & Industry (hereafter DTI), 2005; Equal Opportunities Commission (hereafter EOC, 2004)). In terms of A levels, the qualifications usually taken at the end of the first two-year period of non-compulsory education (16-18 years), girls also proportionately achieve more A-C passes in these subjects (EOC, 2004).

Achievements in computing courses have followed this trend. Over the last decade, girls have gained around 40% of GCSE qualifications in computer studies, and they are now more far likely to gain an A-C pass than their male counterparts (EOC, 1994-2004). Nevertheless, at A level, when students traditionally specialise in three or four subjects, the trend has been for the majority of girls to opt out of computing. In 1993, in England, girls only accounted for 13% of students deciding to further their computing education to A level standard in England (EOC, 1994). By 2003, this picture had significantly improved, with girls comprising 26% of those studying computing or information technology A level (Joint Council for Qualifications, 2004). Although this still represents a substantial “leak” between one educational level and its successor, it is noteworthy that girls have recently become proportionately more likely to gain the top grades in these qualifications as well (DTI 2005; Joint Council for Qualifications, 2004).

COMPUTER SCIENCE AT THE HIGHER EDUCATION LEVEL

As Figure 1 indicates, the proportion of women within computer science courses at tertiary levels remained fairly static between 1994-2003, despite the improving proportion of them taking computer science at A Level over the same period. Although there appears to be a slight increase between 2002 and 2003, this is likely to be due to changes in the way graduate statistics in the UK were calculated between these two years. On average, women comprised 22% of those completing a degree in the area over the period.

In the UK, female applicants to computer science courses differ in key respects from their
male counterparts. They are especially likely to cite their interest in the subject as the main reason for studying it and are committed to finding the right course above other considerations, such as location of university and so forth (Connor, Burton, Pearson, Pollard, & Regan, 1999; Craig, Galpin, Paradis, & Turner, 2002b; Millar & Jagger, 2001). This interest would seem to be intrinsic, as they are less likely to cite future employment prospects as a motivating factor for their subject choice than their male counterparts (Millar & Jagger, 2001). Female computer science applicants have indicated that, in order to aid course choice, they review a wide range of information and are especially influenced by the information provided in university prospectuses and by experiences of pre-entry campus visits. Indeed, information deriving from universities is deemed far more helpful than advice from secondary school teachers and careers advisers (Connor et al., 1999; Craig et al., 2002b; Millar & Jagger, 2001).

Women are slightly more likely to be accepted for a degree within the broad area of mathematical sciences and Informatics, of which computer science is a subset, probably due to their better success rates at A-levels. They are also more likely to complete their degrees (DTI, 2005).

Once on their degree courses, although they are now over-represented within higher education as a whole (comprising 56% of new graduates), as Figure 1 has illustrated, women remain significantly outnumbered within each cohort of computer science undergraduates. This is in a context where computer science is attracting more undergraduates year-on-year whereas physics, chemistry, and engineering are attracting fewer, and mathematical sciences broadly the same numbers (DTI, 2005). As Figure 2 chronicles, between 1994/5-2002/3 the numbers of men and women completing degrees in this area in the UK climbed throughout the period, with the percentage of men climbing slightly faster than the percentage of women. In 1995, 5.5% of all male graduates qualified in computer sciences, as against only 1.5% of all females. By 2003, 11.8% of males were graduating in this area, as against 2.8% of females. While there are grounds for optimism insofar as the proportionate increases during this period for women were comparable to those of men, it remains of concern that while more than a tenth of male graduates take their degrees in this area, less than 3% of females do. This is significant in the context of the link between computer science and professional-level IT work, and the critical role computer science knowledge plays in our society more generally.

Women are also generally under-represented among the academic faculty of computer science at university, although their numbers in Lecturer (28%) and Researcher (30%) ranks remains better than we might expect given their undergraduate representation (DTI, 2005). They are, however, disproportionately underrepresented in the more senior academic ranks (DTI, 2005).

Interestingly, however, as Figure 3 shows, of those taking degrees in this area, women fare no worse than men in terms of classification achieved,
and their performance would seem to be improving. In 1994/5, 9.6% of men achieved the top grade of degree—a First Class—as opposed to 7.6% of women. By 2003, however, 11.9% of men were achieving this grade against 11.6% of men. Furthermore, women have also been marginally, but consistently, more likely to obtain a “good” degree—a First Class or Upper Second—than men since 1994, and by 2003 just over 50% of them achieved this standard as against just under 47% of men.

**FUTURE TRENDS AND CONCLUSION**

Once this picture is digested, we can see that there are some grounds for cautious optimism about the future: an increase in female participation at A level; improvements in female performance in relation to both pre-university and university qualifications; and evidence of above-average interest in computer science on the part of female undergraduates. Notwithstanding these changes, a key question is begged, however, given the context of the overall increase in the importance of the subject: Why is the under-representation of women that established itself in the field in the 1980s and 1990s not correcting itself more quickly?

It is still claimed by some that women are naturally less predisposed than their male counterparts to be attracted to, or to perform well within, the sciences in general (McCrum, 1996) and computer science in particular (Craig et al., 2002b). This position, however, is increasingly difficult to sustain in the light of changes in female participation and performance rates in educational computing, and the fact that there are significant local variations in the uptake and experience of university-level computer science courses (Cahoon, 2001; Craig et al., 2002b; Millar & Jagger, 2001). As Cahoon has suggested, it is much more likely that social factors are at play to explain female under-representation as “women can and do succeed in Computer Science when conditions do not deter them” (2001, p. 1).

These “social factors” would seem from the evidence to fall under two main areas. First, females, including female computer science undergraduates, project a far more negative image of IT work than males. As their educational choices are so closely bound up with their projected career trajectories (EOC, 2001; Miller, Neathey, Pollard, & Hill, 2004), it is reasonable to assume that this tendency has impacted on female recruitment levels to the subject at university. Secondly, female experiences of computer science degrees would seem to be less positive than those of their male counterparts, and it is very likely that the stories of such experiences become part of the information young women draw on when making their decisions about what subjects to study at university.

To take the first point, the available evidence suggests that women taking computer science degrees, and courses that might prepare them for

---

**Figure 3.**

<table>
<thead>
<tr>
<th>Year</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1985</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Based on original analysis of Higher Education Statistics Agency data provided under ad hoc data enquiry: 23148
undergraduate study, are generally not likely to be fully aware of the benefits of a career in IT (Millar & Jagger, 2001; Peters, Lane, Rees, & Samuels, 2003), while simultaneously being more likely to be alive to its real and perceived disadvantages. They are more likely to project, for instance, that it will be solitary work, when they, as a group, have a preference for team-based and socially-embedded computing (Craig et al., 2002b). The problems facing them should they opt for a career in computing cannot be ignored, substantial as they are. Female computer science graduates are less likely than males to find employment at the end of their degrees, despite their tendency to have achieved a higher degree classification, and to be better educated more generally. Once in employment, they are likely to receive less pay for comparable work, and many describe a working ethos that systematically privileges men over women (Millar & Jagger, 2001; Woodfield, 2000). In a context where female educational achievements are outstripping those of males at every level of education in the UK, and female work aspirations are becoming increasingly clear and uncompromising, it is almost certainly the case that the negative perception of IT work plays a critical role in determining the numbers of women choosing courses that qualify them for it. As Millar and Jagger have claimed, problems appear in the relationship between women and the computer sciences “once women begin to structure their education towards a career” (2001, p. A-iv). There are substantial benefits to IT work, however, that could be clarified and communicated to those selecting their GCSE, A level and university courses: salary levels, notwithstanding the gender gap, tend to be higher than average; existing female IT professionals rarely describe themselves as bored or unfulfilled in terms of the work per se. More fundamentally than the need to communicate these benefits, the sector needs to consolidate the beneficial aspects of IT careers for women, and work to eradicate the more negative aspects.

The earlier communication about the advantages of IT work takes place, the better. It is claimed that educational choices, based on projected career costs and benefits, are already taking place when UK teenagers face their first opportunity to specialise, around age 14. This is also an age when young males and females are most sensitive to gender-appropriate and gender-inappropriate career choices. It is of concern then, that there seems to be so few reports of positive representations of professional IT work as a viable career goal for girls (and their parents) by teachers or careers advisers within secondary schools, and that the provision of female-friendly spaces for computing is still considered cutting edge (Craig et al., 2002b; Millar & Jagger, 2001; Peters et al., 2003). Improvements in the knowledge, and sometimes the attitude, of adults playing a key role in pre-university course choices, could make a significant difference to female participation rates at this level.

As well as the role played by the image of IT work, the reported experiences of those undertaking a computer science degree must also partially account for the level of female participation at university. There is a wealth of evidence to suggest that female students’ under-representation on courses creates its own problems in terms of their confidence, adjustment, enjoyment, and achievement levels. They report themselves to feel pressurised to adapt to a male-dominated and male-oriented educational regime in order to survive (see, for example: Bjorkman, Christoff, Palm, & Vallin, 1998; Blum, 2001; Margolis & Fisher, 2002; Peters et al., 2003). Evidence from those institutions that have attempted to make significant changes within their computer science programmes so that women will feel less alienated suggests that the have enjoyed far greater female participation and approval rates as a result (Connor et al., 1999; Margolis & Fisher, 2002; Millar & Jagger, 2001). In doing so, they point to a wish list of best practices for others interested in doing the same. These include building a curriculum that
assumes a minimal level of previous IT experience, providing mentoring systems and female-friendly learning spaces, and ensuring that faculty are sensitive to the needs and abilities of female students (Blum, 2001; Cahoon, 2001; Hefce 2002; Margolis & Fisher, 2002). One course in the UK, where such considerations have been central for some time, reports high rates of female uptake of degree places, but also that 100% of women on the course maintained a desire to become an IT professional after the completion of their degree (Craig et al., 2002).

The finding that women are heavily reliant on the information provided in prospectuses is heartening in this regard. It confirms that there is a clear window of opportunity for the more progressive universities through which they can encourage potential applicants and reduce reliance on general anecdotal or impressionistic information that may not apply to their courses. If more institutions embraced the same ethos, a critical mass of women entering computer science undergraduate programmes could be achieved in the next decade. This, in turn, could impact significantly on the general culture of educational computing and beyond into the workplace, which could lead to improvements in the sector’s image. The beneficiaries of this would not just be the women who would otherwise have turned away from the subject at degree level, but also the UK economy and society which is arguably at present drawing its computing scientists and skilled IT professionals from an artificially restricted pool.

REFERENCES


KEY TERMS

Classification of Degree: UK degrees are ascribed “classifications” that usually correspond to the following percentage results (averaged over all assessments contributing to the final grade):

- First – 70%+
- Upper Second or 2:1 – 60-70%
- Lower Second or 2:2 – 50-60%
- Third – 40-50%
- Pass – 30-40% (a bachelor degree without honours)

Equal Opportunities Commission/EOC: The Equal Opportunities Commission is an independent, non-departmental public body, funded primarily by the government. It provides free research reports and statistics chronicling the position of women and men in all aspects of UK life, including work and education. Web address: http://www.eoc.org.uk

GCE A Level: The General Certificate of Education, A level (Advanced Level) is usually taken 2 years after GCSEs and as a non-compulsory, pre-entry qualification for University-level Study. Students usually take 3 or 4 specialised subjects at this level, which they are free to select.

GCSE: The General Certificate of Secondary Education is usually taken by 15-16 year olds in a range of subjects—some compulsory (English, maths, and science) and some optional (including Computer Studies)—at the end of their compulsory schooling. Achieving 5 GCSE grades A-C is considered the desirable educational benchmark to be reached by this age.

Good Degree: A term used in academic literature and common parlance to indicate a degree which has been classified as an “Upper Second” or “First,” (i.e., in the top two degree classifications. It is normally assumed that those going on to postgraduate study will have such a
degree, and they are preferred in many employment contexts.

**HESA (Higher Educational Statistics Agency):** HESA was set up in 1993 by the UK Government to act as a central source for higher education statistics and has become a respected point of reference. Web address: http://www.hesa.ac.uk.

**IT/Computer Professional:** An individual working within the IT/computer sector within a complex and skilled role that is classified within the category “Professional occupations” (e.g., software engineers), or sometimes within “Associate professional and technical occupations” (e.g., computer programmers)—in the National Statistics classifications, Standard Occupational Classification Codes (2004): http://www.statistics.gov.uk

---

**ENDNOTES**

1. Elsewhere, degrees in mathematical sciences and/or business and administration degrees are included in analyses of gender representation within degrees associated with IT and IT employment (see, for instance, Millar & Jagger, 2001). Here, unless stated otherwise, I will be using the narrower group of “computer science” degrees, defined, according to Higher Educational Statistics Agency programme categories as including: computer science, information systems, software engineering, artificial intelligence, and other programmes within computer science.

2. All of the statistics included within this section are based on original analysis of Higher Education Statistics Agency data provided under *ad hoc data enquiry: 23148*.


This work was previously published in the Encyclopedia of Gender and Information Technology, edited by E. Trauth, pp. 365-371, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 4.11
Gender and Information Technology in Rural Bangladesh

Lutfor Rahman
Association for Advancement of Information Technology (AAIT), Bangladesh

Nusrat Rahman
University of Dhaka and Association for Advancement of Information Technology (AAIT), Bangladesh

INTRODUCTION

Information technology (IT) is transforming our personal, social, cultural, economic and political lives. But women in developing countries do not have equal access to knowledge, due to the fact that they do not have access to the new technologies at the same level as western European women. They need to understand the significance of new technologies and use them in relevant fields. IT can offer immense opportunities for virtually all girls and women in developing countries, including poor women living in rural areas.

Developing countries like Bangladesh are usually seen as problematic hosts for ITs because most developing regions of the country lack economic resources and indigenous techno-scientific capabilities to develop and deploy IT infrastructures. The regions also tend not to make the best use of opportunities of technology transfer.

The wider adoption of IT in the 1990s promised a globally connected community of equal participants in electronic networks. The Internet has enabled people to communicate regardless of race, sex, physical ability, location and social background (GVU, 1998). IT now offers access to a huge pool of information. The Internet enables worldwide communication that is cheaper, faster and more flexible than older media like e-mail, telephone, telex or fax.

BACKGROUND

In early 2000, an initiative was taken to develop a small river island (char) of Bangladesh through
establishment of a resource centre named Indigenous Science and Technology at Ikrail (ISTI), and using IT as the main vehicle for development. IT is a livelihood to many women of the developed countries but is almost unknown to the women and girls of the river island.

Women in Bangladesh are seen in the frontlines to fight against hunger, poverty and environmental degradation. Because of lack of awareness about the benefits of IT, they cannot raise their problems for solution to the proper authority. IT can benefit women in various ways and can facilitate their participation in different sectors and regions. It can provide information women need to improve their own well-being and that of their families. The introduction of computers in offices can improve the quality of work and scope for women in data entry, analysis, programming, clerical and administrative occupations in Bangladesh. IT could allow them to exchange views, opinions and information with women of other regions and countries.

The situation for rural populations in many regions of the world is characterized by geographical distance from urban centres that offer employment and income, education, cultural events and public services. IT bears the potential to improve the situation of rural people in several ways. For example, it is agreed that early diagnosis of medical issues can prevent many casualties. Once a patient comes to the health centre, using various sensors, information can be collected and communicated to an expert at hospitals in the district headquarters. The primary healthcare centre could be located in a rural area or in a mobile centre installed in a van.

IT offers access to information and extension, to education and training communication and networking in online discussions. IT also offers access to employment and income.

**Infrastructural Problems**

Poor healthcare can result from a lack of good information. Establishing reliable communications may be one of the most important priorities for improving healthcare and education. Many rural areas in the globe have no or outdated telephone lines that can transmit Internet-based data. The lack of infrastructure can be overcome by wireless technology (e.g., radio modems). Mobile telecentres can be a solution for the target group.

**HEALTHCARE IT PROJECT**

A significant number of women scientists, researchers and technologists work in rural Bangladesh. They are disadvantaged and traditionally underrepresented in most respects. Their knowledge and skills are unrecognized, underused and under valued. As such, they are in greater need of upgrading their skills, especially in the fast advancing world of information and communication technologies (ICTs), which might enable them to connect their global colleagues, sources of information and global knowledge wherever they may be located.

A survey conducted in early 2000 among 515 women professionals of various disciplines spanning life sciences, social sciences, physical sciences, mineral sciences, engineering, technologies and medical sciences identified that they are almost illiterate in IT, although they are well qualified and experienced in their respective fields. At the first step, the survey was conducted over the senior professors of the public and private universities, scientists of the Bangladesh Council of Scientific and Industrial Research (BCSIR), researchers of institutes of forest research, housing research, fuel research, jute research and water resource and geological research.
They have very poor knowledge about the power and benefit of IT. This trend was found true at all levels of educational backgrounds. The survey revealed that the illiteracy rate in IT among the medical doctors are high, because the doctors are normally resistant to new technology, particularly IT.

To reach the benefit to the rural communities, participants for IT training were selected from among applicants spanning all sectors, giving priority to the rural community. The representation covered various national economy, especially health, environment, education, natural resources (mineral resources), energy, industry, agriculture, fisheries and forestry. Special care was taken in selection of participants so that the maximum number of organizations and individuals would benefit from the program. Training programs on IT for the duration of 2 weeks were organized in different private universities at a certain interval with 20 participants in each phase. The programs were initially organized at the capital city, and later, the programs were arranged at the regional institutes and universities. The regional programs were mainly organized for the medical doctors working in the local hospitals, clinics, medical colleges and universities. Selection of participants from the women medical doctors was really a tough job. Neither the medical doctors nor their family members wanted to spare 2 weeks for attending a training course that they feel are good for their children, not for the doctors. Strategies were adopted to attract the medical doctors. As soon as they realized the importance of IT, they invested more of their valuable time to learn further applications. Ninety-two percent of the trained medical doctors were found to utilize their gained knowledge intelligently in their respective fields. The result was obtained from the follow-up visit by the project staff.

**FUTURE TRENDS**

Bangladesh is a country of rivers that create islands (chars) of different shapes and sizes. They are created because of river erosion. The ISTI resource centre is located in one such island. The island is circular in shape and its area varies from 20 to 30 square kilometers, depending on seasons. Approximately 15,000 people live in the island community. The people depend on local cultivation for survival. Their average monthly earning is $40 (United States dollars). The area is fertile and rice, wheat, peanut, jute, sugarcane, and different types of vegetables and fruits are common. All work is performed manually. The people are deprived of basic infrastructure. There is no electricity, phone system or sanitation. The only source of water is contaminated with arsenic. There are no health clinics or doctors in the area. Many adults are illiterate, and curable diseases, such as cholera, malaria and tuberculosis, are pervasive.

The ISTI resource centre has been established with the objective of developing this isolated region with proper education, empowering local people with relevant technology, particularly IT, and providing them with medical facilities using modern techniques (such as telemedicine and telehealth care). The ISTI resource centre, equipped with basic medical test equipment like stethoscopes, blood pressure meters and blood testing chemicals, and connected with a hospital located in the district town, might save lives of many women and girls. Just a computer with a modem connected with wireless local loop (WLL) help solve most of the serious problems of local women by the women medical doctors empowered with IT. The ISTI resource centre is working as role model for several million disadvantaged women living in the river islands of Bangladesh.
CONCLUSION

This training program has been particularly suitable and effective for relatively marginalized groups and remote communities in a developing country because of its low cost and the use of local organizers and trainers, courses especially designed to meet the participants’ needs and in situ follow up. The direct involvement in the courses of senior policy makers, employers and local experts has proven to be absolutely crucial in gaining support for the continuation and expansion of the course and, perhaps more importantly, effecting attitude change to women’s roles and capabilities in science and medicine. This was underlined by the increasing cooperation extended by them to the project team during preparation and execution of the project. The outcome of the whole training has been impressive. Participants’ value to the work place has increased, some attained promotions and some changed careers.

The final evaluation report shows that most women medical doctors and other health-related professionals empowered with IT have expressed their opinions that every medical doctor must attend relevant technology—particularly IT—courses if they want to enhance their services and maintain their personal secrecies. Some mentioned that IT is a magic and all-pervasive tool.

REFERENCES


KEY TERMS

Information and Communication Technology (ICT): ICT covers computing, electronics and telecommunications.

Medical Professionals: Medical doctors and scientists in health-related subjects, such as nutrition, bio-chemistry, psychology and so forth.

Researchers: Graduates and post graduates engaged in research activities.

Scientists: Science graduate with some research experience in scientific fields.

Technologists: Engineers and scientists with experience in hardware.

University: A university is characterized by a wide variety of teaching and research, especially at a higher level, that maintains, advances, disseminates and assists the application of knowledge, develops intellectual independence and promotes community learning.
Chapter 4.12
Chinese POS Disambiguation and Unknown Word Guessing with Lexicalized HMMs

Guohong Fu
The University of Hong Kong, Hong Kong

Kang-Kwong Luke
The University of Hong Kong, Hong Kong

ABSTRACT

This article presents a lexicalized HMM-based approach to Chinese part-of-speech (POS) disambiguation and unknown word guessing (UWG). In order to explore word-internal morphological features for Chinese POS tagging, four types of pattern tags are defined to indicate the way lexicon words are used in a segmented sentence. Such patterns are combined further with POS tags. Thus, Chinese POS disambiguation and UWG can be unified as a single task of assigning each known word to input a proper hybrid tag. Furthermore, a uniformly lexicalized HMM-based tagger also is developed to perform this task, which can incorporate both internal word-formation patterns and surrounding contextual information for Chinese POS tagging under the framework of HMMs. Experiments on the Peking University Corpus indicate that the tagging precision can be improved with efficiency by the proposed approach.

INTRODUCTION

While a number of successful part-of-speech (POS) tagging systems have been reported for English and many other languages over the past years, it is still a challenge to develop a practical Chinese POS tagger due to the language-specific issues in Chinese POS tagging. First, there is not a strict one-to-one correspondence for a Chinese word between its POS and its function in a sentence. Second, an ambiguous Chinese word can act as different POS categories in different contexts without changing its form. Third, there are many unknown words in real Chinese text whose POS categories are not defined in the dictionary.
Chinese POS Disambiguation and Unknown Word Guessing with Lexicalized HMMs

used. Furthermore, there are almost no explicit morphological features in Chinese words, such as inflexion, derivation, and capitalization in English, that can be used directly for POS disambiguation and unknown word guessing (UWG). All these factors make it much more difficult to achieve a high-performance POS tagger for Chinese.

Recent study on POS tagging has focused on machine-learning approaches, including hidden Markov models (HMMs) (Brants, 2000; Weichedel et al., 1993), transformation-based error-driven learning (TBL) (Brill, 1995), maximum entropy model (Ratnaparkhi, 1996), support vector machines (SVMs) (Nakagawa et al., 2001), and loglinear models (Fu & Wang, 2002). Machine-learning approaches have the advantage of robustness. However, it is difficult for most machine-learning techniques to keep a balance between their capacity and their computational cost (Nakawa et al., 2002). Pla and Molina (2004) have proved that HMM-based taggers have the highest training and tagging speed. However, they usually achieve relatively low tagging precision, because standard HMMs only take into account contextual POS information for tagging. On the contrary, some learning models such as ME and SVMs are capable of handling much richer lexical information for POS tagging. However, they usually require higher computational cost, which inevitably will result in reduction of efficiency in training and tagging. This will be a serious problem in processing a large amount of data or in some online applications. In order to tackle this problem, some recent work, such as Lee, et al. (2000) and Pla and Molina (2004) suggested the use of lexicalization techniques to enhance the standard HMMs. Their experiments have shown that the lexicalized models can improve POS tagging precision without increasing much computational cost in training and processing.

POS disambiguation and unknown word guessing are two key issues in developing a Chinese POS tagger for practical applications. On the one hand, Chinese is highly ambiguous with respect to part of speech. Consequently, the first task of a POS tagger is how to find a proper POS for each ambiguous word in a sentence. On the other hand, there are many unknown or out-of-vocabulary (OOV) words in real Chinese texts whose POS categories are not defined in advance in the system dictionary. So, a practical tagger should be capable of predicting or guessing with accuracy the POS categories for these unknown words in an open-ended text.

Following the line of Lee et al. (2000) and Pla and Molina (2004), we propose in this article a unified approach to Chinese POS disambiguation and unknown word guessing. In order to explore word-internal morphological features for Chinese POS tagging, we introduce four types of pattern tags that indicate the way of a lexicon word to present itself in a real segmented sentence. Further, we define a hybrid tag set by merging these patterns with POS tags, with which Chinese POS disambiguation and unknown word guessing can be unified as a single process of assigning each known word (KW) in input a proper hybrid tag. Moreover, a statistical tagger is developed based on the uniformly lexicalized HMMs. In this way, three kinds of features — contextual tag information, surrounding lexical information, and word-internal word-formation patterns — can be incorporated for Chinese POS tagging under an efficient HMM-based framework. As a consequence, the tagger’s performance can be improved with efficiency in training and tagging.

The rest of this article is organized as follows: The second section presents a novel formulation for Chinese POS tagging. Next a uniform lexicalization technique is introduced to enhance the standard HMMs for POS tagging. The fourth section gives a brief description of the tagging algorithm. Finally, we report our experimental results on the Peking University corpus in the fifth section and give our concluding remarks on this work in the final section.
**UNIFYING POS DISAMBIGUATION AND UNKNOWN WORD GUESSING**

In this section, Chinese POS disambiguation and unknown word guessing is reformulated as one unified process of known word tagging, based on the introduction of word formation patterns.

**Word Formation Patterns**

The input of our current task is a segmented sentence or a sequence of words that might be either known words listed in the system lexicon or unknown words unlisted in the system lexicon. In general, a known word consists of only one lexicon word, while an unknown word is composed of several lexicon words. As discussed in Fu and Luke (2005), a lexicon word has four possible patterns to form a word in a segmented sentence: (1) $w$ is segmented as an independent known word; (2) $w$ is the beginning component of an unknown word; (3) $w$ is a middle component of an unknown word; (4) $w$ is the ending component of an unknown word. In this article, four different tags — ISW, BOW, MOW and EOW — are used to denote these patterns, respectively. Thus, a segmented sentence can be represented as a sequence of lexicon words accompanied with their word-formation pattern tags. For example, an equivalent expression of the POS-tagged sentence 张晓华在工作 (Zhang Xiao-hua is working) is 张/ISW晓/ISW华/MOW在/EOW工作/ISW.

**POS Disambiguation and Unknown Word Guessing as Known Word Tagging**

Given a real word in a Chinese sentence, the position of its component lexicon words is another important evidence that can be used to determine its POS in addition to the contextual information surrounding it. For example, if a word ends with the character 性 (xing4, nature), then it is more likely to be a noun. On the contrary, if a word ends with the character 化 (hua4, change or a suffix such as -ize and -ify), then it should be tagged as a verb in most cases. In order to explore these important position features for POS disambiguation and unknown word guessing, a hybrid tag set is thus designed by merging the above four patterns with POS tags. In our work, a hybrid tag has the following format: T1-T2. Where T1 denotes a POS tag and T2 denotes a word-formation pattern tag.

With this hybrid tag set, an original POS-tagged sentence can be represented equivalently as a sequence of known words accompanied by their relevant hybrid tags. For example, an equivalent expression of the POS-tagged sentence 张/nr晓华/nr在/p香港/ns大学/n工作/v张晓华/nr在/p香港/ns大学/n工作/v is 张/ISW晓/ISW华/MOW在/EOW在/p-ISW在/p-ISW在/p-ISW大学/n-ISW工作/v-ISW.

Obviously, the POS category of an unknown word in a sentence will be determined once each of its components is assigned a proper hybrid tag. At this point, the two subtasks in Chinese POS tagging (i.e., POS disambiguation and unknown word guessing) thus can be unified as a single task of tagging each known word in input, with a proper hybrid tag indicating its POS category and its word-formation pattern.

**LEXICALIZED HMMs**

**Uniformly Lexicalized HMMs**

Given a sequence of known words $W = w_1w_2...w_n$, a statistical tagger aims to find an appropriate sequence of hybrid tags $\hat{T} = t_1t_2...t_n$ that maximizes the conditional probability $P(\hat{T}|W)$; namely,
\[
\hat{T} = \arg \max_T P(T \mid W) = \arg \max_T \frac{P(W \mid T)P(T)}{P(W)}
\]  

(1)

where the probability \( P(W) \) can be dropped in that it is fixed for a specific word sequence \( W \). Thus, we have a general model for POS tagging as follows:

\[
\hat{T} = \arg \max_T P(W \mid T)P(T) = \arg \max_T P(w_{1:n}, t_{1:n})
\]

\[
= \arg \max_T \prod_{i=1}^{n} P(w_i \mid w_{i-1}, t_{i-1})P(t_i \mid w_{i-1}, t_{i-1})
\]

(2)

In theory, the model in Equation (2) can provide the tagging system with more powerful capacity for disambiguation and unknown word guessing. However, it involves too many parameters to be computable in practice. To address this problem, two models—the standard HMM and the lexicalized HMM—are usually employed to approximate the general model in Equation (2).

Actually, the standard hidden Markov modeling follows two well-known independent hypotheses: The appearance of current word \( w_i \) depends only on current tag \( t_i \) during known word tagging, and the assignment of current tag \( t_i \) depends only on its \( N \) previous tags \( t_{i-N}, \ldots, t_{i-1} \). Thus, Equation (2) can be rewritten as

\[
\hat{T} = \arg \max_T \prod_{i=1}^{n} P(w_i \mid t_i)P(t_i \mid t_{i-N}, \ldots, t_{i-1})
\]

(3)

Equation (3) is a general form of the \( N \)-order standard HMMs, where \( P(w_i \mid t_i) \) denotes the so-called lexical probability, and \( P(t_i \mid t_{i-N}, \ldots, t_{i-1}) \) denotes the contextual tag probability. Due to the serious data sparseness in real applications, only the first-order or second-order models are employed in most current POS tagging systems (i.e., \( N = 1 \) or \( N = 2 \)).

At present, standard HMMs have been used widely in many NLP applications such as automatic speech recognition (ASR) and POS tagging, due to their efficiency in training and processing. Furthermore, a variety of theories and algorithms have been proposed for building HMMs with efficiency. However, a standard HMM-based tagger only can take into account contextual category information for tagging. The important contextual lexical features often are ignored. As a consequence, HMM-based taggers usually have relatively low tagging performance. In order to enhance tagging performance and give consideration to efficiency in training and tagging at the same time, we employ the uniform lexicalization technique (Lee et al., 2000) to enrich the standard HMMs with contextual lexical information.

In approximating the general model in Equation (2), the uniformly lexicalized modeling adopts two assumptions: the appearance of current word \( w_i \) is assumed to depend not only on current tag \( t_i \) but also its previous \( N \) words \( w_{i-N}, \ldots, w_{i-1} \) and tags \( t_{i-N}, \ldots, t_{i-1} \), and the assignment of current tag \( t_i \) is supposed to depend on both its previous \( N \) words \( w_{i-N}, \ldots, w_{i-1} \) and the relevant \( N \) tags \( t_{i-N}, \ldots, t_{i-1} \). Thus, a uniformly lexicalized HMM can be formulated as

\[
\hat{T} = \arg \max_T \prod_{i=1}^{n} P(w_i \mid w_{i-N}, \ldots, w_{i-1}, t_{i-N}, \ldots, t_{i-1})P(t_i \mid w_{i-N}, \ldots, w_{i-1}, t_{i-N}, \ldots, t_{i-1})
\]

(4)

In comparison with standard HMMs, uniformly lexicalized HMMs can handle both contextual words and contextual tags for the assignment of hybrid tags to known words, which will result in an improvement of tagging precision. In view of the serious data sparseness in higher-order models, we employ the first-order lexicalized HMMs in our system.
Estimation and Data Smoothing

If a large POS-tagged corpus is available, the lexicalized HMMs easily can be estimated using the maximum likelihood estimation (MLE) technique. In principle, MLE estimates parameters using their relative frequencies that can be directly extracted from the training data. Obviously, MLE has the advantage of simplicity. However, it will yield zero probabilities for some cases that are not observed in the training data. Relatively speaking, a lexicalized HMM-based tagger will be faced with a more serious problem of data sparseness than a standard HMM-based tagger. In order to tackle the problem of zero probabilities in MLE, we employ the linear interpolation smoothing technique to smooth the lexicalized parameters using the relevant non-lexicalized probabilities. This process can be formulated with Equation (5):

\[
\begin{align*}
F'(w_1 | _w_{-1}, \lambda) &= \lambda F(w_1 | _w_{-1}, \lambda) + (1 - \lambda) F(w_1 | _\lambda) \\
F'(\lambda | _w_{-1}, \mu) &= \mu F(\lambda | _w_{-1}, \mu) + (1 - \mu) F(\lambda | _\lambda)
\end{align*}
\]

(5)

where, $\lambda$ and $\mu$ refer to the smoothing coefficients.

THE ALGORITHM

This section describes an algorithm for Chinese POS tagging. The problem of inconsistent tagging also is discussed in this section.

Viterbi Tagging

Given a sequence of words, there might be more than one possible sequence of POS tags. The task of a tagging algorithm is to find the best one that has the highest score according to the models in Equation (3) or (4). In our system, we apply the classical Viterbi algorithm to perform this task, which works in three main phases as follows:

1. **Decomposing of unknown words.** As discussed previously, lexicon words (i.e., known words) are taken as the basic tagging units. In this phase, all unknown words in input are decomposed to a sequence of known words using the forward maximum match (FMM) method (Liang, 1987), and each decomposed known word then is assigned a proper pattern tag in terms of its position in the relevant unknown word. At the same time, each known word in the input is labeled a pattern tag ISW.

2. **Generation of candidate POS tags.** In this phase, all possible POS tags first are generated for each word in the input: For a known word, its POS candidates are generated by looking up the system lexicon; For an unknown word, its POS candidates are created by searching the lexical probability library in terms of its components and their word-formation patterns. In accordance with the format defined in the third section, these POS candidates are then merged with the relevant pattern tags created in phase (1). All these combined candidate tags are stored in a lattice structure.

3. **Decoding of the best tag sequence.** In this step, the Viterbi algorithm is employed to score all candidate tags using the language models given in the fourth section and then to search the best path through the candidate lattice that maximizes the score. This path contains the most probable sequence of POS tags for the input. It should be noted that the direct output of our tagger is represented as a sequence of known words together with a sequence of hybrid tags. For evaluation purposes, such a format will be converted further into the normal tagging format by
removing the pattern tags and reconstructing
unknown words using their components.

**Inconsistent Tagging**

Inconsistent tagging will arise when the components of an unknown word are labelled with different POS tags. For example, the personal name 晓华 in the sentence 张晓华在香港大学工作 (Zhang Xiaohua is working at the University of Hong Kong) might be inconsistently tagged as 晓/Vg-MOW 华/nr-EOW. In this case, the system will be unable to make its decision in determining the POS category of the full word 晓华. Consequently, how to avoid inconsistent tagging is a fundamental issue that relates to the practicability of our system.

To achieve a complete consistent tagging, we develop a rule-based module in our system, which can prevent inconsistent tags from entering into the lattice for storing candidates during the generation of potential POS tags for unknown words. Given an unknown word and its components, our system first creates a set of POS candidates for each of its components in terms of its relevant pattern and its lexical probability. Then, it will continue to generate the POS candidate set for the unknown word using one of the following four rules: (1) if the intersection of all the POS candidate sets for its components is not null, then the intersection is taken as its POS candidate set; or (2) if the intersection of the two candidate sets for its beginning and ending components is not null, then the intersection is its POS candidate set; or (3) if the union of the two sets of POS candidates for its beginning and ending components is not null, then the union is its POS candidate set; and (4) if the three previous rules do not work, then the top frequent POS tag (i.e., noun) is taken as its POS candidates. It should be noted that once the POS candidate set for the unknown word is determined, all its components should share this set as their POS candidates. Rule (2) stresses the importance of the beginning component and the ending component, because according to our intuition, the two components might be more informative than other components in determining the POS category of a Chinese unknown word in most cases.

**EXPERIMENTS AND DISCUSSIONS**

In order to evaluate the effectiveness of our system, we conducted a number of experiments using the Peking University corpus and lexicon. This section reports the results of these experiments.

**The Experimental Data**

The Peking University (PKU) corpus (Yu et al., 2003) contains six months (January to June, 1998) of Chinese newspaper text from the People’s Daily. As shown in Table 1, the PKU corpus has been segmented manually and contains approximately 7.3 million words. Each word in this corpus also has been tagged with one of 45 POS tag in Appendix A. In our experiments, the first month of data (i.e., January 1998) was used for testing, and the remaining five were for training. It should be noted that an original POS-tagged corpus should be converted to the format shown in the second section before using is for training (i.e., a sequence of known words together with their hybrid tags). In addition to the PKU corpus, we employed a POS dictionary of about 75,000 entries in our system, which were mainly from the Grammatical Knowledge-Base of Contemporary Chinese (Yu et al., 1998). In order to make this lexicon complete, a number of non-Hanzi characters and GBK Hanzi were also added to it.

Table 1 also gives a survey of Chinese POS ambiguities and unknown words in the PKU corpus. It can be observed that over 55% of known words are ambiguous in POS, and some 7% of words in the PFR corpus are unseen in the
dictionary used in our system. This shows that high-performance disambiguation and unknown word guessing are very important for a practical Chinese POS tagging system.

Table 2 shows the top 10 frequent types of unknown words in the PKU corpus. It can be observed that noun is the most frequent type of unknown words. We can see that more than 20% of unknown words are nouns in the PKU corpus. As can be seen from Table 2, the top 10 frequent types account for about 93.80% of unknown words in the training corpus and 94.20% of unknown words in the testing corpus.

The Experimental Results

In our evaluation, we conducted two experiments on the PKU corpus in Table 1:

1. Since a lexicalized HMM-based approach can handle richer contextual information for tagging, particularly the contextual lexi-
Table 3 shows the results of our first experiment. We can see that the lexicalized HMMs perform better than the standard HMMs as a whole. In comparison with the standard HMM-based tagger, the lexicalized HMM-based tagger can improve the tagging precision respectively by 2.06% for all words, 1.88% for all known words, 3.37% for ambiguous known words, and 4.42% for unknown words. At the same time, the use of lexicalized models does not lead to a rapid increase of tagging time. This indicates that the lexicalized HMM-based method is able to keep a relative balance between tagging precision and tagging efficiency.

Figures 1, 2, and 3 present the curves of the tagging precision for all words, ambiguous known words, and unknown words, respectively, vs. the size of the data for training. We can see that in comparison with the standard HMM-based tagger, the tagging precision of the lexicalized HMM-based tagger is changing in a sharper ascent curve for all three cases as the size of the training corpus increases, which indicates that the uniformly lexicalized HMMs need much more training data than the corresponding non-lexicalized models in order to achieve a reliable estimation. In other words, the performance of a lexicalized HMM-based tagging system is more sensitive to the size of the training data than that of a standard HMM-based tagging system. Furthermore, it can be observed from Figure 2 and Figure 3 that a larger improvement of tagging precision can be achieved for unknown word

<table>
<thead>
<tr>
<th>Method</th>
<th>Tagging accuracy (%)</th>
<th>Tagging speed (w/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>KWs</td>
</tr>
<tr>
<td>Standard HMMs</td>
<td>94.08</td>
<td>94.52</td>
</tr>
<tr>
<td>Lexicalised HMMs</td>
<td>96.14</td>
<td>96.40</td>
</tr>
</tbody>
</table>
Figure 1. Overall tagging precision vs. the size of training data

Figure 2. Precision for Amb. KWs vs. the size of training data
guessing than for the disambiguation of known words. To our knowledge, the reason might be that unknown words in real text form an open set, while known words are a close set. In general, more training data will provide the system with much richer contextual information and word-formation cues for the guessing of different unknown words in an open-ended text. As a consequence, an increase of training data will result in a steady improvement of tagging precision in unknown word guessing.

CONCLUSION

We presented a unified resolution to Chinese POS disambiguation and unknown word guessing. In order to explore word-internal cues for Chinese POS tagging, we introduced four types of word-formation patterns that indicate whether a lexicon word is an independent word or a component word in a specific position of an unknown word in a sentence. We further defined a hybrid tag set by merging these patterns with POS tags, with which we can reformulate Chinese POS disambiguation and unknown word guessing as one single process of assigning each known word by inputting a proper hybrid tag. To do this, we developed a statistical tagger based on the uniformly lexicalized HMMs. In this way, three types of features — word-internal word-formation cues, contextual tags, and contextual words — can be incorporated to resolve different ambiguous words or unknown words in Chinese text. We also evaluated our system using the Peking University corpus and dictionary. The experiments have shown that our system can yield correct results for most ambiguous words and unknown words in question. Our experimental results also indicated that the use of the lexicalization technique is able to improve the
tagging precision without losing much efficiency in tagging.

ACKNOWLEDGMENTS

We would like to thank the Institute of Computational Linguistics, the Peking University for their corpus and lexicon. We also would like to thank the two anonymous reviewers for their helpful and valuable comments.

REFERENCES


ENDNOTES

1 Unlike other languages like English, Chinese text has no explicit delimiters to mark word boundaries except for some punctuation marks. Therefore, word segmentation is usually the first step in performing POS tagging on a text in Chinese. An introduction of Chinese word segmentation can be seen in Fu and Luke (2005). This article focuses on Chinese POS tagging.

2 It should be noted that a normal input word sequence might contain known words and unknown words. In this article, each unknown word will be decomposed into a sequence of known words using the efficient forward maximum match technique (Liang, 1987) before POS tagging.
## APPENDIX A.

**The PKU Part-of-Speech Tagset**

<table>
<thead>
<tr>
<th>Tag</th>
<th>Definition</th>
<th>Tag</th>
<th>Definition</th>
<th>Tag</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>形容词</td>
<td>Rg</td>
<td>代语素</td>
<td>nz</td>
<td>专有名词</td>
</tr>
<tr>
<td>ad</td>
<td>副形词</td>
<td>Tg</td>
<td>时语素</td>
<td>o</td>
<td>拟声词</td>
</tr>
<tr>
<td>an</td>
<td>名形词</td>
<td>Ug</td>
<td>助语素</td>
<td>p</td>
<td>介词</td>
</tr>
<tr>
<td>b</td>
<td>区别词</td>
<td>Vg</td>
<td>动语素</td>
<td>q</td>
<td>量词</td>
</tr>
<tr>
<td>c</td>
<td>连词</td>
<td>Yg</td>
<td>语气语素</td>
<td>r</td>
<td>代词</td>
</tr>
<tr>
<td>d</td>
<td>副词</td>
<td>h</td>
<td>前缀</td>
<td>t</td>
<td>时间词</td>
</tr>
<tr>
<td>e</td>
<td>叹词</td>
<td>i</td>
<td>成语</td>
<td>u</td>
<td>助词</td>
</tr>
<tr>
<td>f</td>
<td>方位词</td>
<td>j</td>
<td>简略语</td>
<td>v</td>
<td>动词</td>
</tr>
<tr>
<td>g</td>
<td>语素</td>
<td>k</td>
<td>后缀</td>
<td>Ag</td>
<td>形容词语素</td>
</tr>
<tr>
<td>Bg</td>
<td>区别语素</td>
<td>m</td>
<td>数词</td>
<td>Bg</td>
<td>形容词语素</td>
</tr>
<tr>
<td>Dg</td>
<td>副语素</td>
<td>n</td>
<td>名词</td>
<td>Mg</td>
<td>数语素</td>
</tr>
<tr>
<td>Mv</td>
<td>名语素</td>
<td>nr</td>
<td>人名</td>
<td>Mg</td>
<td>数语素</td>
</tr>
<tr>
<td>Qg</td>
<td>量语素</td>
<td>nt</td>
<td>组织名</td>
<td>N</td>
<td>状态词</td>
</tr>
</tbody>
</table>

*This work was previously published in the International Journal of Technology and Human Interaction, edited by B. C. Stahl, Volume 2, Issue 1, pp. 39-50, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 4.13
IT Implementation in a Developing Country Municipality: A Sociocognitive Analysis

Clive Sanford
Aalborg University, Denmark

Anol Bhattacherjee
University of South Florida, USA

ABSTRACT
This article presents an interpretive analysis of the key problems and challenges to technology implementation in developing countries, based on a three-year case analysis of an IT project in a city government in Ukraine. We employ the concept of technological frames of reference as an analytical tool for articulating the group-level structures related to the implementation context from the perspectives of key stakeholders and examine the degree of conflict between these frames using a Fishbone diagram. We report that conflict between technological frames held by key stakeholders in large-scale system implementation projects often create an unexpected, dysfunctional, and politically charged implementation environment, ultimately leading to project failures, even if the project enjoys a high level of financial and management support. This, in turn, creates unique challenges for technology implementation projects in developing countries that are often overlooked in the traditional academic and practitioner literatures based on experiences from developed countries.

INTRODUCTION
Information technology (IT) has long been viewed by central planners in the developing world as an important tool for achieving rapid economic and wage growth, improving operational efficiency and effectiveness, and enhancing political participation and transparency. However, achievement of these objectives is often thwarted due
to incentive structures that are based on existing rules and organizational structures. Improvements in administrative capacity can only be expected when there is a sound institutional base that is supported by operational, technical, and infrastructural facilities. Therefore, planning for the diffusion of IT into a developing country’s administrative strategies presents a challenge that is significantly different from that encountered by developed countries.

The UN Secretary-General has stated, “information technologies can give developing countries the chance to leapfrog some of the long and painful stages of development that other countries had to go through” (Annan, 2002). In other words, IT has the potential to support the development strategy of bypassing some of the processes of the accumulation of human capabilities and fixed investment in order to narrow the gaps in productivity and output that separate industrialized and developing countries (Steinmueller, 2001). Recently, there is also the indication that the creation of an information society and, in particular, e-government implementation would lead to better (or good) governance (Ciborra & Navarra, 2005). However, the public and businesses in developing countries often have a mistrust of governments that are entrenched in hierarchical structures that are often politicized and corrupt, and are rife with cumbersome processes that stunt economic growth and discourage investment in productive business activities. Efforts to restructure government are either in direct conflict with institutional archetypes or have proven difficult for countries with unstable governments to adopt. In addition, government agencies in developing and transitional countries often try to install IT infrastructures that have been designed for private sector firms and/or for governments in other countries.

Globalization, democratization, and economic liberalization have prompted initiatives with IT as the primary lever to stimulate dramatic changes in the role of the state. New freedoms and opportunities, especially prevalent in developing countries, have raised the expectations of individual citizens, and emerging and increasingly vocal and proactive civil societies now do not hesitate to call governments to account. This is somewhat mitigated by the fact that IT in many public sectors has not delivered the value expected (Bellamy & Taylor, 1994; Lenk, 1990; Margetts, 1999; Margetts & Willcocks, 1993; Willcocks, 1994). The literature on IT in developing countries generally reports on planned systems, pilot studies, and failures, and seldom discusses the outcome of projects that are even partially functional (Cecchini & Raina, 2005). Further, even when systems are operational, sustaining operations in an ongoing manner is difficult (Frasheri, 2003).

This article presents a longitudinal case analysis of an IT initiative in the city government of L’viv, a municipality in western Ukraine, to improve governance and service quality for its citizens. The specific system under investigation is a Lotus Notes based document management system (DMS), intended to automate and streamline citizen services such as building construction permits, business license applications, and zoning clarifications. A similar project was rolled out in the Malaysian government Generic Office Environment (GOE) (Karim, 2003). Grant and Chau (2005) used the Malaysian GOE as a case example of service delivery in their strategic focus areas of their e-government framework. While a governmental DMS has an objective that is congruent with e-government to “enable effective constituent relationship management” (Grant & Chau, 2005, p. 9), and in general may qualify as a specific example of e-government, we prefer to orient the discussion of a DMS as an effort to improve e-administration. E-administrative initiatives deal particularly with projects that involve the re-engineering of internal structures and organizational activities. These initiatives include cutting process costs, managing process performance, making strategic connections in government, and “creating empowerment by transferring power,
authority and resources for processes from their existing locus to new locations” (Heeks, 2002b, p. 99). This requires a fundamental restructuring of the public administration’s organization and knowledge management practices by integrating work processes across departments and agencies in order to build a system that simplifies interactions and improves services.

We describe three years of the deployment history of the DMS project (2002-2005), by drawing on the experiences of one of the study’s authors, who was a participant observer in this project at this time. The author was commissioned by the City Mayor as a consultant to provide user training for the DMS and related systems. Based on numerous personal interactions with people involved at all levels of this project, such as the Mayor’s office, end users, the city’s IT department, and external consultants, and direct involvement with the DMS and prior L’viv city projects, this was a unique opportunity to unravel the complex dynamics of the DMS project as it evolved over time. Our interpretive analysis of interview, observational, and documentary evidence revealed several interesting findings and implications about the politics of IT implementation projects in developing countries. Given the unprecedented nature of political, sociocultural and economic change witnessed in Ukraine and other new democracies of eastern Europe over the last decade, this context provided us an ideal testbed for studying the processes and impacts of e-administrative government projects in countries that are seeking to establish western-style democratic reforms.

The structure of the article proceeds as follows. The next section discusses the rationales behind the introduction of IT in developing countries. The third section describes a sociocognitive view of IT implementation, which serves as the theoretical backdrop for our research. The fourth section presents a background of the context of this case study, followed by a historical description of the events in the case. The fifth section describes our research methods, including data collection and analysis approaches. The sixth section presents our research findings, based on interpretive analysis of the case data. The seventh section provides a discussion of the key issues in the case and the final section proposes directions for future research and presents concluding remarks.

THE DIFFUSION OF IT TO DEVELOPING COUNTRIES

Information technology and associated organizational models, which form the basis of information systems and implementation methodologies, have originated and continue to be developed principally in western, developed countries. The unidirectional transfer of IT knowledge and skills from developed to developing countries is a process that characterizes most recent efforts at public sector reform (Minogue, 2001). However, in order to take advantage of technological knowledge of developed countries, developing countries must have acquired sufficient technological capabilities and institutional capacities to identify suitable technologies and to adapt, absorb, and improve the technologies imported from abroad (Ahrens, 2002). Often, they are not appropriate for use in countries that are in transition and have their own realities on the design of technologies that are used to automate and inform their internal and citizen-focused processes. Fountain (2001) addressed this issue in her differentiation between the deployment of already invented technologies that are available to practitioners and the need to recognize the importance of customizing the design to a specific locale.

A rationale that is often used for the introduction of IT is to automate, increase efficiencies, and cut costs. The cost of labor in developed countries is substantially higher than that in developing countries (Dewan & Kraemer, 2000), while the reverse is true with the cost of the IT (Heeks, 2002b). Therefore, replacing inexpensive civil servants with more expensive IT would not be jus-
tifiable for financial reasons alone. Consequently, the choice of projects and their ultimate success is dependent on careful consideration of the embedded web of relations and interactions within a particular socioeconomic context, and their design and implementation requires an understanding of this context (Sein & Harindranath, 2004).

Governmental and international development agencies (IDAs) in the developed world, such as the World Bank, the United Nations, and United States Agency for International Development (USAID), have spearheaded and provided resources for deploying governmental IT technologies in the developing world. Since nearly half the people of the world today are under 25 years old and 90% of these young people live in developing countries (World Bank, 2005), IDAs are seeking to optimize the deployment of their resources to parts of the world where they would have the potential to have the most enduring and significant impact. Their initiatives fit well with the western ideals of globalization, economic liberation, and democratization, with improved governmental functioning in developing countries helping to secure important sources for raw materials or low-cost labor for western businesses and opening up new or growing existing markets in these countries for western goods and services.

IDAs and others believe that the scarcity of IT in developing countries is one of the reasons for decreased opportunities for economic growth and social development in developing countries (Madon, 2004). However, such initiatives are challenged by the lack of integrated information and communication technology infrastructures, antiquated work practices, and perhaps most importantly, organizational obstacles such as conflicting objectives, lack of widespread support, and internal resistance to change within governmental agencies. Madon (2004) explored an alternative perspective of development where conditions are created to encourage individuals to realize their full potential by freely selecting their path to improvements in material, social, and spiritual lifestyles (Sen, 1999). The World Bank (2005) has encouraged such investments that empower people to participate in decisions about development projects that affect their lives and the lives of their families and have an enduring impact on their quality of life.

A Sociocognitive Perspective of Implementation

The sociocognitive perspective is based on the idea that reality is socially constructed through human beings’ interpretation of social reality and negotiation of meaning among social entities (Berger & Luckmann, 1967). Like the cognitive perspective, the sociocognitive perspective emphasizes the role of one’s personal knowledge structures in shaping her cognitive interpretation of the world and consequent actions. However, it differs from the cognitive perspective in the way it views the formation of individual knowledge structures. The cognitive perspective postulates information processing and learning as the key processes shaping individual knowledge structures, while sociocognitive research holds that shared knowledge and beliefs (group-level structures), imbibed through the processes of interaction, socialization, and negotiation within a larger social group, influence individual sense-making and interpretation. These shared knowledge structures, also called “frames” (Goffman, 1974) or “interpretive schemes” (Giddens, 1984), serve as templates or filters for individual sense-making and problem solving, filling information gaps with information that conforms with existing knowledge structures, and systematically rejecting inconsistent information (Fiske & Taylor, 1984). Frames also provide a vehicle for organizing and interpreting complex and sometimes ambiguous social phenomena, reducing uncertainty under conditions of complexity and change, and justifying social actions.
The sociocognitive perspective has seen some applications in prior information systems research. In the context of information systems development, at least three studies (Bansler & Bodkar, 1993; Davidson, 2002; Newman & Nobel, 1990) have posited that requirements for IT applications do not exist a priori but are socially constructed through interactions among system developers, user groups, and other key stakeholders. The language used by participants and the nature of their interaction influences which requirements are identified and legitimized (Boland, 1978).

Drawing from the above research, Orlikowski and Gash (1994) proposed the concept of technological frames of reference as an analytical tool for understanding and articulating group-level structures that guide sense-making and action. Technological frames are defined as “assumptions, knowledge, and expectations, expressed symbolically through language, visual images, metaphors, and stories” that organizational members employ to interpret the role of technologies in organizations and understand the conditions and consequences of such use (Orlikowski & Gash, 1994, p. 176). Technological frames typically operate in the background, and have facilitating or constraining effects on the interpretations and actions of their intended users (Orlikowski & Gash, 1994).

Given the diversity of stakeholders often involved in large-scale IT implementation projects, such as system developers, senior management, and multiple user groups, it is likely that different stakeholder groups have different technological frames that may be partially or wholly incongruent with each other. For instance, technologists often view IT from an engineering perspective, as a tool specifically designed to accomplish a given task such as system development. In contrast, managers tend to take a more strategic view of IT, focusing on its return on investment and long-term value. End users may take a more instrumental view of the same technology, expecting specific work-related benefits from its use. Orlikowski and Gash (1994) suggest that when incongruent technological frames exist, organizations are likely to experience difficulties and conflicts in implementing IT, including breakdowns in communication, lack of user participation, and/or eventual suspension or failure of the project.

In his pioneering work on force-field analysis, social psychologist Kurt Lewin (1951) postulated that social systems such as organizations continuously seek equilibrium between forces favoring change and those opposing change, and that successful change rests on an organizations’ ability to “unfreeze” the equilibrium by altering the dynamics of forces favoring and resisting change. Without explicit attention to and manipulation of the forces opposing change, organizational systems may tend to resist change and retain the status quo (similar to the concept of “homeostasis” in biological systems).

Lewin’s ideas are applicable to understanding IT implementation because organizational introduction of IT is a social change process since it changes the way organizational members work, interact with each other, and make decisions. Given the natural behavioral resistance to change, technological frames once formed tend to be resistant to change. However, contextual changes, such as changes in key stakeholder groups, changes in resource distribution, or emergence of new emergent threats, can trigger shifts in frames by bringing new ideas, assumptions, and knowledge to the forefront of the group sense-making process, forcing the social group to reinterpret old information in new ways. El Sawy and Pauchant (1988) note that such “frame shifting” may be abrupt and of short duration; nonetheless they can engender long-term changes in the group’s overall stance toward IT and IT-enabled work.

In summary, sociocognitive analysis can complement our traditional cognitive view of IT implementation by bringing into focus the role of group-based structures in individual sense-mak-
IT Implementation in a Developing Country Municipality

ing and interpretation of social phenomena during the implementation process. We pursue such analysis using the concept of technological frames as an analytical lens to further explore the role of social structures during IT implementation.

CASE BACKGROUND AND DESCRIPTION

Given that technological frames are time- and context-dependent (Orlikowski & Gash, 1994), in this section, we first present a detailed picture of the social context within which the target IT was implemented, followed by a description of the actual implementation process.

The Sociopolitics of Ukraine

Ukraine is an eastern European country which became independent from the erstwhile Soviet Union in 1991. Limited exposure to free markets and state protection of enterprises from foreign competition since the Soviet days has led to an antiwestern sentiment among many Ukrainians, suspicion of western economic interests, an entrenched governmental bureaucracy, lack of corporate tax reforms, widespread organized crime, and corruption among many government officials. Transparency International (2004) ranks Ukraine as 122nd out of 145 countries on its corruption index. For instance, local elected officials often use their influence over water, electricity, and sanitation departments to force private enterprises to contribute to social programs and other less-productive ends.

Ukraine’s IT infrastructure is ranked 78th out of 102 countries in the Networked Readiness Index (Dutta & Jain, 2004). A United Nations e-government readiness survey (2004) rated Ukraine a value of 0.53 on a 0-1 scale and an overall rank of 45 out of 191 UN member states. This ranking indicates the willingness and readiness among governments to employ the opportunities offered by IT to improve the access and quality of basic social services for sustainable human development. In 2004, Ukraine improved its ranking by nine positions, but further improvements are clearly needed to reduce poverty, unemployment, state control, and corruption, and to create an environment more conducive for private businesses.

Municipal government is an important element in Ukraine’s public administration system. Each city is headed by a Mayor, who is elected by the entire city electorate. The Mayor is answerable to the municipal council (or city council), whose members are elected by the people to four-year terms. At the start of this project, city council members at Lviv were in office for about a year, with three years remaining in their term of office. The city administration consists of the Mayor and directors and division heads nominated by the Mayor for each of the city’s subdivisions, pending confirmation by the city council.

At the time of the project, Lviv’s information systems (IS) department consisted of 35 employees, responsible for supporting the city’s IT infrastructure and application development and maintenance efforts. The director of this department reported to the Organization and Information Subdivision, which in turn, was accountable to the city council. However, like other city departments, the IS department was also highly compartmentalized with little interaction with other departments or awareness of others’ needs. The IS director was often unaware of what software was used by other subdivisions, and therefore did not support its usage. This lack of coordination sometimes led to idiosyncratic IT implementations that were not coordinated across city departments, that lacked any plan for long-term maintenance, and that were sometimes not even useful.

E-Administration at Lviv City Hall

The first phase of the e-administrative initiatives at Lviv started in 1994 as part of a $3 million
USAID grant (see project timeline in Figure 1). The goal of this grant was to improve citizen service and government accountability by streamlining internal operations in three Ukrainian cities: L’viv (population: 850,000), Kharkiv (population: 1.6 million), and Ternopil (population: 230,000). Though no organizational analysis or needs assessment was conducted at Kharkiv and Ternopil prior to technology deployment, L’viv’s IS deployment did follow a minimal analysis conducted by a local consultant group recruited and funded by USAID. The consulting group’s main office is located in the United States, and operated from a branch office in Kyiv that employed local Ukrainians.

The primary directive of the consultant group was to install hardware and software infrastructures per USAID guidelines, leaving subsequent implementation activities, such as user support and training, and ultimate system usage to the grant recipients. Though technically proficient, the consultant group was not adequately experienced in large, complex process re-engineering initiatives of this magnitude. Hence, the city’s IS department was left out of the initial decisions regarding choice of hardware, software, and networking technology, and how they would be deployed. The internal IS department was assigned a role only after the consultant team left at the end of the year 2000. Disenchanted with their omission in key decision processes regarding the city’s IT infrastructure, the IS department had little interest in supporting the city’s IT initiatives.

In 1998, the consulting group designed a conceptual framework for improving internal workflow within the City Hall to serve as the foundation for a new document management system (DMS), and subsequently used the Lotus Notes software to customize the DMS to the city’s needs. Lotus Notes was selected because of its compatibility with both Macintosh and PC operating systems. An internal audit of the DMS-enabled customer service by the Organization and Information Subdivision on a sample of three city departments during the first half of 2003 revealed some surprising inconsistencies in system use and usage outcomes. Though city departments were supposed to record all correspondence in the DMS, this mandate was not enforced and was widely ignored. Less than 20% of the filed requests were actually entered into the DMS, and the remaining requests were retained

---

**Figure 1. Timeline of Government IT Implementation at L’viv**

<table>
<thead>
<tr>
<th>Event</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>USAID approved government IT grant for three Ukrainian cities</td>
<td>1994</td>
</tr>
<tr>
<td>Consultants commissioned to choose and install IT infrastructure</td>
<td>1995</td>
</tr>
<tr>
<td>Lotus Notes based DMS installed</td>
<td>1998</td>
</tr>
<tr>
<td>City audit showed little IT usage; New project champion</td>
<td>Early 2003</td>
</tr>
<tr>
<td>Consulting left as USAID funds ran out; City takes over</td>
<td>Late 2003</td>
</tr>
<tr>
<td>IT infrastructure upgraded; Universal Office rolled out</td>
<td>2001</td>
</tr>
</tbody>
</table>

---

Phase 1 | Phase 2
in article format for manual processing. Of the 2,405 citizen requests processed by the DMS during this six-month period, only 184 requests (8%) were completed within 30 days as required by city regulations, and an additional 905 requests (38%) were simply “lost in the system” in that no information was available regarding their final disposition (see Table 1).

Concerned with the above situation, the secretary to the city administration initiated new plans to enhance the levels and outcomes of DMS usage at L’viv City Hall. This second phase of IT implementation at L’viv focused more on change management and garnering user support, than on technology selection or installation. The secretary established a one-stop shop called the “Universal Office” at the city correspondence office to serve as a central clearinghouse for all citizen and business requests. The office was staffed with technically proficient people whose job was to enter and track all incoming user requests into the DMS system right at the source. Further, one of this study’s authors was commissioned by the city at this time to provide technology training to the intended users of the DMS.

As of early 2005 (the end of the case analysis period), the status of DMS usage at L’viv was as follows. Most of about 8,300 citizen service monthly requests received by the City of L’viv were still being hand-delivered or mailed to the appropriate city department by citizens instead of being filed at the Universal Office as originally envisioned. The preferred methods of communication among citizens and city employees still remained telephone and face-to-face meetings. About 25% of city employees used the DMS’s e-mail feature on a regular basis, and another 50% used it occasionally. The DMS was solely used by administrators to track documents, and other document lifecycle features such as version histories and audit trails were not being widely used. Each city department employed a staff member to handle internal correspondence electronically and exchange internal documents and files through a shared file system. Further, no interdepartmental correspondence was being tracked through the Notes system at the time.

DATA COLLECTION AND ANALYSIS

Case studies are particularly suitable for answering “how” and “why” questions, when multiple sources of evidence are needed, and when boundaries between the observable facts and the context are not clearly evident (Yin, 1994).

Table 1. Six-month document volume and dispositions

<table>
<thead>
<tr>
<th>Department</th>
<th>On time (%)</th>
<th>Violation (%)*</th>
<th>No information (%)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing</td>
<td>175 (21)</td>
<td>442 (53)</td>
<td>217 (26)</td>
<td>834</td>
</tr>
<tr>
<td>Architecture</td>
<td>8 (1)</td>
<td>563 (73)</td>
<td>200 (26)</td>
<td>771</td>
</tr>
<tr>
<td>Natural Resources</td>
<td>24 (3)</td>
<td>288 (36)</td>
<td>488 (61)</td>
<td>800</td>
</tr>
<tr>
<td>Total</td>
<td>184 (8)</td>
<td>1293 (54)</td>
<td>905 (38)</td>
<td>2405</td>
</tr>
</tbody>
</table>

* A document is in violation if it was either not answered or answered incorrectly.
It also enables researchers to use “controlled opportunism” to respond flexibly to new discoveries made while collecting new data (Eisenhardt, 1989). Such was the nature of this study, where the individual usage of frames was evident in the discourse that was related to IT requirements as well as written artifacts (Moch & Fields, 1985). In addition, a longitudinal case study design is desirable for observation and analysis of processes that change over time (Glick, Huber, Miller, Doty & Sutcliffe, 1990).

We therefore followed Walsham (1993) and Avgerou and Walsham (2000) in applying Pettigrew’s (1985) methodology for examining the embedded web of relations and interactions between the choice of the IT and the context of its use. Thus, the research was conducted as an interpretive case study of the innovation provided by the technology as perceived by the IDA and other stakeholders where the primary purpose was to better understand the complexities of the subjective meanings of human subjects within cultural and contextual situations usually found in field studies (Orlikowski & Baroudi, 1991). Since in interpretive research the data and subsequent analysis are influenced by the researcher’s own preconceptions and inquiry processes (Walsham, 1995, p. 376), the use of technology frames concept guided interpretation of the subjects’ actions and the events in the DMS project. Throughout the study, key informants listed in Table 2 were interviewed and informal discussions about DMS project events were held.

Typical interviews ranged between 20 and 90 minutes, with an average duration of about 45 minutes. During the interviews, open-ended questions were asked with the intent of understanding the missions, goals, visions, and strategies being used, the expected and actual organizational impacts, and the influence of cultural factors in gaining acceptance and usage of the technology. In addition to the face-to-face personal interviews, nine group meetings with both city and departmental administrators were conducted at multiple stages of the project between 2002 and 2005. Given the sensitivity of government personnel and consultants to privacy issues, interviews were not tape recorded. However, the interviewer maintained extensive field notes for each interview, including actual quotes from interviewed personnel, which were used for subsequent data analysis.

Several additional forms of documentation were also employed for data collection. These included the city’s vision statement and planning documents, internal memos, working papers, PowerPoint presentations, and USAID reports.

### Table 2. Informants interviewed

<table>
<thead>
<tr>
<th>Role in DMS Project</th>
<th>Number of interviews</th>
<th>Role in DMS Project</th>
<th>Number of interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>City Mayor</td>
<td>2</td>
<td>Departmental administrators</td>
<td>4</td>
</tr>
<tr>
<td>Secretary to the city administration</td>
<td>4</td>
<td>Staff personnel</td>
<td>4</td>
</tr>
<tr>
<td>Secretary to the city council</td>
<td>4</td>
<td>IS department administrator</td>
<td>5</td>
</tr>
<tr>
<td>Directors of subdivisions</td>
<td>5</td>
<td>Consultant</td>
<td>2</td>
</tr>
<tr>
<td>Informants interviewed</td>
<td>8</td>
<td>Total interviews</td>
<td>30</td>
</tr>
</tbody>
</table>
Further, the researcher observed city employees at work at the city correspondence office and within individual departments, participated in several IT planning meetings, and had numerous informal interactions with key participants that were involved in the DMS project for over two years. This direct involvement provided a holistic understanding of the complex, multifaceted project as well as the individual behaviors, comments, and perspectives of the involved parties, by iterating between the parts and the whole in a “hermeneutic circle” (Klein & Myers, 1999).

The first step in our sociocognitive analysis was to identify the key frame domains. Orlikowski and Gash (1994) asserted that frames should be examined in situ, rather than assumed a priori, via an iterative examination of observed data and experiences. Hence, we analyzed our qualitative data to identify the assumptions, knowledge, or expectations (frames) of the DMS technology implicit in participants’ direct or symbolic words (e.g., metaphors) or actions, and the implications of these frames for participants’ work practices and City Hall’s operations as a whole. Our qualitative data were first separated into groups based on whether it reflected the statements of end users (e.g., project managers or city administrators), executive management (e.g., Mayor’s office, secretary of city administration, etc.), internal (city) IS department, and external consultants. Field notes and interview quotes for each group were coded into frame categories as suggested by the data using an open coding style typical of grounded theory research (Strauss, 1987). Following this exercise, we conducted cross-group analysis, where we compared the categories generated by each group to elicit common themes between groups, as recommended by Eisenhardt (1989).

The iterative examination of data described above resulted in the identification of four common themes or domains that are described here as technological frames. These domains were then examined for similarities and differences between different stakeholder groups. Higher dissonance between these frames was expected to lead to greater inconsistency and incoherence between activities by multiple stakeholder groups, thereby increasing the chances of project failure. The findings of our analysis are described next.

RESEARCH FINDINGS

Four dominant technological frames or common themes, representing the salient perceptions, assumptions, and interpretations of diverse stakeholder groups, emerged from our interpretive analysis of qualitative data from L’viv’s DMS implementation project:

1. **Purpose of technology:** Stakeholders’ perceptions of organizational motivation for implementing a specific IT and the reasonableness of such motivation.
2. **Technology features/capabilities:** Stakeholders’ understanding of the IT’s design, features, and functionalities, whether or not such perceptions were justified or realistic.
3. **Technology implementation process:** Stakeholders’ perceptions of how the IT implementation process was managed, including such issues as user involvement, user training, and change management, and whether or not such implementation is reasonable or effective.
4. **Technology-induced changes:** Stakeholders’ expectations of how IT implementation may change existing work practices, relationships, and responsibilities, and whether or not such changes are justified.

In the following subsections, we describe each of our four proposed frames from the perspectives of five key stakeholder groups involved in L’viv’s DMS implementation project: the sponsor (USAID), the city management (Mayor’s office and the city council), the external consultants,
the internal IS department, and the targeted users (department administrators and staff members).

**Purpose of Technology**

**Sponsor’s Perspective**

The sponsoring agency (USAID) had envisioned a western view of government functions, stressing the importance of citizen service quality, efficient organizational processes, and modern technologies. Government functions, such as business license processing or city zoning clarifications, were viewed as information processing activities that involved complex patterns of information sharing and collaboration among multiple governmental agencies. Such information processing activities were expected to benefit from the deployment and utilization of advanced IT such as a DMS. Additionally, automated e-administration was expected to indirectly increase the transparency of government functioning, enhance public accountability of city functions, and reduce corruption in city administration.

**Consultants’ Perspective**

The external consultants did not play a significant role in shaping the purpose or motivation of IT implementation at City Hall. Their mandate was simply to take USAID’s espoused goals as given and select and deploy a technological infrastructure (hardware, software, and network) within cost and schedule guidelines as specified by the sponsor, similar to the western norms of IT project management.

**City Management’s Perspective**

IT initiatives at L’viv started in 1994, well before the current city administration was elected to office, and hence, the Mayor’s office was not involved in early stages of the project. Initial work on the project was sponsored by USAID, with the support of the then city council. Post-Soviet western influence suggested that building appropriate IT systems were the best way to support cross-departmental data flow at the city hall for faster and accurate processing of customer requests. As stated by the secretary to the city council:

*Our departments need to respond to citizen-centric views that involve cross agency data and processes, and they will need to structure their systems, data and programs to be able to deliver on this objective.*

Additionally, city council members also wanted to portray a pro-technology and pro-reform image of themselves to their electorate in order to enhance their chances of re-election success. However, without strong leadership from the mayor and incentives to actually use the DMS, department heads were not motivated to use a system that they viewed to be a threat to their power. It was primarily the non-utilitarian considerations of citizen perceptions of the city council that amplified the city’s motivation to deploy technologies that support e-administration.

**IS Department’s Perspective**

Even though the city’s IS department had expertise in areas such as needs analysis, technology identification, and technology deployment, it had no role during the initial phase of the project. The department was therefore disenfranchised from the city council’s vision of improving their image to the electorate. Perhaps more importantly, they viewed the introduction of the technology to be incongruent with current workflows and were unprepared to assume responsibility for training city hall staff on its proper usage.

**User Department’s Perspective**

The elected officials of the city had a shared goal of improving government functionalities and
citizen service quality, but the administrators and staff personnel at the city departments did not share similar objectives. One administrator felt that the DMS system did not fit the face-to-face communication needs of his job and was more of an intrusion on his work, rather than benefiting it, though other IT systems such as office productivity software were more useful:

*If the mayor tells us to use the system then I guess we will have to, but I really don’t like it. Computers just get in the way of solving problems in face-to-face meetings. I don’t think that I would want to trust my staff to follow my decisions through the system. These are always matters that I should be handling.*

On rare instances, however, a small number of city administrators appeared to agree on the potential benefits of the DMS system and expressed some enthusiasm toward it. One of them commented:

*This notes system is really interesting. It looks like it will help me do my work faster and make better decisions. With this kind of system it will be easy for me to find answers to my problems.*

**Technology Features/Capabilities**

**Sponsor’s Perspective**

USAID’s goal was to improve government transparency and citizen service through the use of technology. However, being an IDA rather than a technology specialist, USAID was not familiar with different governmental solutions, their specific features or capabilities, or the technological infrastructure required to implement such solutions. Hence, it left technology selection and implementation decisions to the external consultants, who were familiar with the specific technologies as well as USAID policies and procedures for technology deployment in governmental agencies.

**Consultants’ Perspective**

Based on their prior involvement with USAID projects (at other municipalities and cities) and their understanding of western procedures of IT implementation projects (e.g., tracking project costs and schedules), the consultants used a least-cost and shortest-route IT implementation process. Lotus Notes was deployed as the workflow system of choice because of its support for e-mail and file sharing capability, and its history of successful applications in document management tasks in western countries.

**City Management’s Perspective**

The L’viv city administration was not involved in the technology selection process during early stages of the e-administration initiative, though they had approved the deployment of the technology at City Hall under the USAID grant. The city’s administrators became more involved in the project only after the USAID grant expired and the consultants left, when the onus fell on the city to maintain and expand the infrastructure and technologies already installed.

**IS Department’s Perspective**

The city’s internal IS staff were not considered technology experts by other city departments, and were not consulted during key technology acquisition or deployment decisions. Hence, their opinions were of no consequence to the sponsor, the implementers (consultants), or the city, who planned to implement the DMS without the IS department’s assistance. Since they were not participants in either the selection or installation of the technology, they were ambivalent about the technology changes both during and after
IT Implementation in a Developing Country Municipality

deployment, and were not willing to actively engage in its implementation until instructed to do so by the city council.

User Departments’ Perspective

Many of the user departments did not have the technology background or expertise to understand, appreciate, or critique the design elements of the implemented technologies. They viewed IT implementation as a technical process left best to domain experts such as the consultants.

Technology Implementation Process

Sponsor’s Perspective

As noted earlier, USAID maintained a “hands-off” approach to the project. They entrusted the consultants with the task of planning and managing the implementation process.

City Management’s Perspective

Like the sponsor, the city management also left the responsibility of managing initial stages of the project to the consultants. When USAID funding expired and the consultant team left, the city was left with complete ownership of the project and was forced to decide how to implement the rest of the project, upgrade necessary systems and the infrastructure, and enhance system usage among users.

Consultants’ Perspective

The consultants did not understand or see the need for user support, process re-engineering, and change management that often accompany large IT implementation projects. They also failed to engage key stakeholders in the project, namely user departments and the city’s IS department.

IS Department’s Perspective

Excluded from key decisions involving the early planning and deployment of the IT, the city’s IS department did not feel any ownership or obligation toward the project. They always felt that the project was “someone else’s problem,” and did not want to get involved during the second phase of the project when invited to do so by the mayor’s office.

User Departments’ Perspective

Most city administrators and staff did not know much about the DMS, its purpose, or how it could benefit their work. Barring a few exceptions, most city administrators did not feel that the system was necessary to perform their job, and in some instances, when asked to do so, simply passed on the responsibility to their subordinates or secretaries. One departmental administrator commented:

_They do not pay me enough for me to spend time learning computer skills. I will continue to do my job the same as always. Besides, nobody uses the Notes product anyway._

Technology-Induced Changes

Sponsor’s Perspective

Though USAID viewed its role as only a funding agency and was not party to any implementation-related decisions, the fact that it allocated no resources for user training or any other change management process suggests that it may have viewed IT-induced changes that often accompany large IT projects as being less relevant or important, or a process that is best relegated to the user organization (City Hall).
City Management’s Perspective

The city management was largely unaware of the complexities in large IT implementation projects and did not proactively plan for change management initiatives at the start of the project. However, when the low state of system use and usage outcomes were revealed during the 2003 audit, the city administration was forced to take notice. The secretary to the city administration, who had some prior exposure to large projects, commissioned an external trainer (one of the paper’s authors) to train City Hall users. He also instituted a series of steps, such as expanding the Universal Office staff, to try to motivate DMS use among city employees.

Consultants’ Perspective

Like the sponsors, the project consultants were not concerned with organizational changes induced by IT or the need to proactively plan for such changes. They were satisfied with the installation of the IT infrastructure and wrote glowing reports about themselves to USAID. They felt that their role in the project ended with system installation, and that they had fulfilled their objective and satisfied USAID’s project requirements. When informed about problems of non-use and city administrators’ resistance to DMS use in 2003, the director of the consulting group responded:

*L’viv has a successfully integrated system of municipal information management which is sustainable at present and is unique for Ukraine. We have no more funding for this project and cannot justify any further need to assess the ICT infrastructure of City Hall.*

IS Department’s Perspective

The IS department’s initial resentment toward the project and its key stakeholders continued into the implementation period, even after the departure of USAID and outside consultants. They were not convinced of the utility of the DMS system and associated organizational changes instituted by the city secretary, and were not unhappy to see it head toward failure. The IS director summarily rejected the idea of the Universal Office as:

*This is a bad idea and will never work. Changing the way citizens submit their requests and training new employees on how to process citizen requests will be impossible.*

User Departments’ Perspective

City Hall users demonstrated a general apathy toward the system and associated changes in workflows and organizational procedures, which were likely caused by low governmental wages and little opportunity for job growth or promotion. Most were subservient to their supervisors and were not interested in engaging in any critical thinking. They viewed the nature and substance of their tasks and activities at City Hall as being largely inconsequential, and were only intent in logging their work hours into the system to satisfy their supervisor. These users had no interest in any technology-induced changes that could potentially increase their workload, even if it promised improved service quality to the citizenry.

DISCUSSION

The sociocognitive analysis used in this study revealed the presence of dominant technological frames that existed prior to the project and persisted well after the installation and organizational implementation of the DMS system. In our study, we found four such frames—purpose of technology, technology features/capabilities, technology implementation process, and technology-induced changes—that were relevant to the context of DMS implementation at L’viv City Hall. Given the generic nature of these frames,
and their conceptual overlap with similar frames reported in prior research (e.g., Davidson, 2002; Orlikowski & Gash, 1994), we can reasonably infer that these frames are not specific to government projects or to developing countries only, but may be generalizable to many large-scale IT implementation projects at large.

Technological frames act as a “sociocognitive lens” through which key stakeholders in an IT implementation project view and interpret reality, and thereby, also condition their responses and behaviors related to implementation processes and outcomes. Some of these frames may be based on pre-existing worldviews or biases that are extraneous to the project context, such as a typical Ukrainian distrust for western interests, lack of a culture of customer service, and limited opportunities of job growth and career enhancements within governmental bureaucracies. Nonetheless, such extraneous mindsets still play a powerful role in enabling or constraining their perceptions of and behaviors associated with IT implementation projects. Other frames may emerge as key stakeholders observe the process and/or outcomes of the IT implementation project. For instance, the IS department, despite being one of the stronger advocates of modern technologies at Lviv City Hall, decided to oppose the DMS implementation project because they were excluded from the planning and implementation phases of the project. In other words, they adopted a political stance to the project, in contrast to the economic stance of improved customer service, faster processing times, and better document tracking espoused by the project sponsors.

What does the notion of technological frames imply for the process and outcome of IT implementation projects? If the frames of key stakeholders are somewhat convergent with each other, then the implementation process is likely to proceed more smoothly than if the frames were divergent to begin with. There may be challenges and difficulties faced during the implementation process, but the shared sociocognitive frame of key stakeholder groups will ensure the formulation of a coherent strategy that would address most implementation-related challenges and collective action to overcome these challenges. On the other hand, dissonant frames among key stakeholders will limit the amount of consensus building, shared interest in the project outcome, and collective action, thereby hindering the implementation process and outcomes. In the case of IT implementation at Lviv, the latter was clearly the case, which ensured an unfavorable project outcome despite financial sponsorship by an external agency and top management support from the city administration. In other words, large-scale IT implementation projects may fail despite financial and management support, if it engenders conflicts in dominant technological frames held by the major stakeholders. This assertion is consistent with prior academic studies on IT project failures such as Lyytinen and Hirschheim (1987), who contended that organizational politics is a key reason for IT project failures, and practitioner studies such as Heeks (2002a), who observed that government information systems projects often appear to be destined for failure.

We depict the conflict in technological frames in the Lviv case pictorially using a Fishbone diagram in Figure 2. This diagram has four branches, each dedicated to one of the four technological frames underscored in this study. In a sense, each branch represents the “force” of each key stakeholder group on the technological frames in this implementation project. Lewin (1951) posited that such “force-fields” depict a dynamic gestalt psychological environment created by multiple driving and restraining forces in the phenomenon of interest. The driving forces are those frames that positively impact the project’s goals of enhancing citizen service quality, document processing efficiency, and transparent government functioning. Stakeholders’ influence on these forces are indicated in our Fishbone diagram as arrows pointing to the right. Restraining forces act to restrain or inhibit the driving forces, either by negatively influenc-
As shown in the Fishbone diagram, the project sponsor and City Hall management had positive influences on the purpose of technology in the IT project by establishing the mission and goals of the project and by securing an external grant to financially support the project. The consultant, management, and sponsor had positive influences on the technology features/capability of the DMS system in that they viewed the system as one that could contribute to productive ends and improve City Hall functioning. However, the IS department and user departments had a negative perception of the technology’s capability, as evident from their remarks about the technology and the futility of its implementation. As stated before, such negative views may have been caused by their pre-existing biases against western agencies (the sponsor) or by virtue of their exclusion from key phases of the project. Nevertheless, this was a negative force that adversely influenced the project outcome. Only the management had a positive impression of how the technology implementation process was managed, while the remaining four stakeholder groups (sponsor, consultants, IS department, and user departments) had either no involvement or were negatively predisposed toward the implementation process. In particular, the disenfranchised IS department refused to accept the implementation process as legitimate, and the absence of support from the sponsor and consultant group during later stages of the project hurt its long-term sustainability, even with the overt support of officials in the mayor’s office. Finally, the technology-induced changes were viewed as positive by the management, but negative by the IS department and user departments. The user
departments refused to accept any changes in their existing work processes or the transparency of their work, despite the technology’s promise of improved process performance.

The end result of the project was that the million-dollar USAID-sponsored governmental IT project at L’viv City Hall that promised to make government processing more efficient and transparent and improve the quality of citizen services, was headed for failure despite financial sponsorship and overt top executive support for the same. This failure was caused by the lack of buy-in and support of the key stakeholder groups during different stages of the implementation process. While it may be tempting to conclude that such failure may occur only in developing countries or with technologically-deficient governments, similar outcomes may also be observed for nongovernmental IT implementation projects (e.g., in universities, health care organizations) in developed countries as well.

What are the key lessons of this IT project? First, financial sponsorship and top executive support are necessary but not sufficient conditions for success of large-scale IT implementation projects. Indeed, IT projects may fail despite the above conditions if they fail to garner adequate support and buy-in from other key stakeholders. Second, the economic or rational perspective of technologies as a tool for improving organizational efficiency, service quality, and transparency may not be adequate for garnering the support of key stakeholder groups. Large IT implementation projects, or most organizational change initiatives for that matter, are often more dominated by sociopolitical considerations than economic considerations (Ewusi-Mensah & Przasnyski, 1991). Hence, mere communication of tangible benefits to stakeholder groups may be inadequate for obtaining their buy-in, and serious negotiations to understand and address issues of ownership, involvement, and power bases may be needed to enhance the chances of IT implementation success. Third, successful implementation requires intra-organizational policy coordination and implementation across stakeholder groups and organizational departments. Alienation or elimination of key departments such as information systems or circumventing key change management initiatives such as user training in order to expedite project goals, may eventually backfire and contribute to the ultimate demise of the project.

In addition to the conditions listed above, government IT implementation initiatives in developing countries such as Ukraine should devote adequate effort to understanding and overcoming sociocultural, legal, political, or other barriers to implementation. For instance, online documents were not officially recognized by the Ukrainian finance and tax system and only hard copy correspondence was recognized as admissible evidence in Ukrainian courts. Such statutes caused many City Hall employees to avoid electronic handling of business documents entirely, thus contributing to system non-usage. Until such restrictive rules were remedied, say by modifying appropriate legal statutes or governmental policies, electronic document management is likely to fail. Currently, a federal law endorsing the use of digital signatures as an acceptable alternative to signed paper documents is pending endorsement by L’viv’s city government, which, when ratified may help reduce public concerns about electronic documents. In retrospect, it therefore appears that the DMS system at L’viv was implemented ahead of its time, and should have been considered after the passage of relevant city laws.

CONCLUSION

The interpretive power of technology frames becomes more apparent when dominant frames become dependent on the active involvement of influential individuals or groups during project activities. This perspective has been addressed through the lens of analyses of power in IT
IT Implementation in a Developing Country Municipality

development projects (Brown, 1998; Markus, 1983). Future research that could further our understanding of the interpretive power of framing would involve the use of actor-network theory (cf., Latour, 1986; Law & Callon, 1988) concepts in order to provide a sense of power, direction, and causality. This would develop our understanding of how stakeholders influence others in their technology frames and thus engender more enduring support for their own interpretation (Lea, O'Shea & Fung, 1995).

The sociocognitive perspective used in this article supports the conclusion that the greatest constraints to e-administrative initiatives are nontechnical, such as political opposition, deeply ingrained policies and practices, and internal employee resistance. To overcome these constraints, governments require proactive, knowledgeable leaders who have an informed long-term vision of technology-induced organization change, who can help spur bureaucratic action, and who can implement strategies that promote and sustain change. In addition, implementers need to not only educate and motivate key stakeholder groups to use the systems, but also identify and redress political inequities and imbalances that hinder e-administrative acceptance among these groups.

REFERENCES


Willcocks, L. (1994, Spring). Managing information systems in UK public administration:
IT Implementation in a Developing Country Municipality

Issues and prospects. Public Administration, 72, 13-32.


Chapter 4.14
Personalization Issues for Science Museum Web Sites and E-Learning

Silvia Filippini-Fantoni
*The University of Paris I Sorbonne University, France*

Jonathan P. Bowen
*London South Bank University, UK*

Teresa Numerico
*London South Bank University, UK*

**ABSTRACT**

E-learning has the potential to be a very personalized experience and can be tailored to the individual involved. So far, science museums have yet to tap into this potential to any great extent, partly due to the relative newness of the technology involved and partly due to the expense. This chapter covers some of the speculative efforts that may improve the situation for the future, including the SAGRES project and the Ingenious Web site, among other examples. It is hoped that this will be helpful to science museums and centers that are considering the addition of personalization features to their own Web site. Currently, Web site personalization should be used with caution, but larger organizations should be considering the potential if they have not already started to do so.

**BACKGROUND**

In the past few years, the number of people visiting museums’ Web sites has gone up rapidly. As a consequence, museums have to face the significant challenge of creating virtual environments that are progressively more adapted towards the different needs, interests and expectations of their heterogeneous users. Increasingly, museums and science centers are using their Web sites to augment their learning facilities in potentially innovative ways.
Personalization Issues for Science Museum Web Sites and E-Learning

(Tan et al., 2003). In particular, museums need to provide for differing online requirements such as teaching, e-learning and research (Hamma, 2004). One of the solutions available to help is the introduction of personalization techniques (Dolog & Sintek, 2004) that, by providing differentiated access to information and services according to the user’s profile, make facilities and applications more relevant and useful for individual users, thus improving the overall visitor’s experience. Science museums, by their very technological nature, ought to be at the vanguard of applying new techniques like personalization.

Developed in the early 1990s in an attempt to try to respond to the different needs and characteristics of an ever-growing number of Internet users, personalized or adaptive Web systems have since been exploited in different sectors such as commerce, tourism, education, finance, culture and health. What distinguishes these systems from the traditional static Web is the creation of a user model that represents the characteristics of the user, utilizing them in the creation of content and presentations adapted to different individuals (Brusilovsky & Maybury, 2002). By so doing, personalization becomes a useful tool in the selection and filtering of information for the user, facilitating navigation and increasing the speed of access as well as the likelihood that the user’s search is successful.

The techniques available to collect information about users, as well as the methods used to process such information to create user profiles and to provide adapted information, are varied. A brief description of the different approaches will be presented here before moving on to illustrate different application examples within the science museum world.

PERSONALIZATION TECHNIQUES

A first important distinction concerning the amount of control the user has on the adaptation process can be made between customization and personalization. Customization or adaptability occurs when “the user can configure an interface and create a profile manually, adding and removing elements in the profile” (Bonnet, 2002). The control of the look and/or content of the site are explicit and user-driven; that is, the user is involved actively in the process and has direct control. In personalization or adaptivity, on the other hand, the user is seen as being passive, or at least somewhat less in control (Bonnet, 2002). Modifications concerning the content or even the structure of a Web site are performed automatically by the system based on information concerning the user stored in the so-called user profile. Such information about the user is provided either explicitly, by the user themselves, using online registration forms, questionnaires and reviewing (static profiles) or implicitly by recording the navigational behavior and/or preferences of each user through dynamic profiling Web technologies such as cookies and Web server log files (Eirinaki & Vazirgiannis, 2003).

Once the data concerning the users is collected either implicitly or explicitly, or even in both ways, as is often the case, appropriate information that matches the users’ need is determined and delivered. This process usually follows one or more of the following techniques: content-based filtering, collaborative filtering, rule-based filtering and Web usage mining.

Content-based systems track user behavior and preferences, recommending items that are similar to those that users liked in the past (Eirinaki & Vazirgiannis, 2003). Collaborative filtering compares a user’s tastes with those of others in order to develop a picture of like-minded people. The choice of material is then based on the assumption that this particular user will value information that like-minded people also enjoyed (Bonnet, 2002). The user’s tastes are either inferred from their previous actions or else measured directly by asking the user to rate products. Another common technique is rule-based filtering, which allows
Web site administrators to specify rules based on static or dynamic profiles that are then used to affect the information served to a particular user (Mobascher et al., 2000).

Last but not least, there is Web usage mining, which relies on the application of statistical and data-mining methods based on the Web server log data, resulting in a set of useful patterns that indicate users’ navigational behaviors. The patterns discovered are then used to provide personalized information to users based on their navigational activity (Eirinaki & Vazirgiannis, 2003).

The information provided to the user through any of the above techniques can be adapted at three different levels: content, navigation and presentation (Brusilowsky & Nejdl, 2004). Adaptive content selection is based mostly on adaptive information retrieval techniques: “when the user searches for relevant information the system can adaptively select and prioritize the most relevant items” (Brusilowsky & Nejdl, 2004). By doing so, the user can obtain results that are more suitable for their knowledge capabilities. Adaptive navigation support is founded mainly on browsing-based access to information: “when the user navigates from one item to the other the system can manipulate the links to guide the user adaptively to most relevant information items” (Brusilowsky & Nejdl, 2004).

Finally, adaptive presentation is based on adaptive explanation and adaptive presence, which were largely developed in the context of intelligent systems: “when the user gets to a particular page the system can present its content adaptively” (Brusilowsky & Nejdl, 2004). The possibilities of content and presentation adaptability are a relevant element in the reuse of the same resources for different purpose, provided they have been correctly customized in advance. Considering the high cost of personalization, adaptability of resources can also offer an interesting byproduct in term of reuse of the same resources in different contexts, provided that their description is correctly defined through standard metadata applications to allow interoperability of the same service in different environments.

From the perspective of different platform services, adaptability becomes a strategic issue. It could be decided to personalize content for the relatively small screen of mobile devices, for example. Moreover, whereas personalization and adaptability on the Web is based only on the user, in the case of mobile support there is also the need for adaptation with regard to the user’s environment (Brusilowsky & Nejdl, 2004).

In a museum visit, taking into account the environment where the service will be used can make a notable difference to the experience. For example, an explanation of the items kept in a single room of the exhibition can be offered while the visitor is in that room. There are some projects exploring these opportunities with special regard to mobile devices used by museum learning services (Oppermann & Specht, 1999).

**WHY USE PERSONALIZATION IN MUSEUMS?**

Even if some of the techniques described in the previous section, especially the more sophisticated ones, are employed mainly on commercial Web sites, such as Amazon.com, etc., there is already some awareness of the need for their use in cultural institutions, museums, science centers, etc. Personalized access to collections, alerts, agendas, tour proposals and audio guides are just a few examples of the different applications that have recently been developed by museums all over the world (Bowen & Filippini-Fantoni, 2004). The reasons for such an affirmation are numerous, as personalization can help museums respond to various and different needs.

First of all, personalization has the advantage of improving the usability of a Web site by facilitating its navigation and aiding people in finding the desired information. With some knowledge about the user, the system can give specific guid-
Personalization Issues for Science Museum Web Sites and E-Learning

in its navigation, limiting the visitation space appropriately. The system can supply, or even just suggest, the most important links or content that could be relevant for the user, something that can help prevent them from becoming lost in a Web site’s potentially intricate hyperspace.

Accessibility for the disabled (Bowen, 2004), a specific aspect of usability that concentrates in widening the number of users, can gain from personalization techniques. The ability to select the text foreground or background color, size and font, can make interfaces more easily readable for the partially sighted. A text only view of a Web site may be easier for such users and also those who are completely blind. For example, the London Science Museum has an option from the home page for a text only version of the Web site (www.sciencemuseum.org.uk). The basic content is the same, but the presentation is different. Legislation in the UK, for example, now ensures that learning materials for students in educational establishments, including those provided by university science museums, must be covered by an accessibility strategy (HMSO, 2001).

Personalized systems help to recreate the human element that listens to the visitor with understanding by offering an individual touch; this is another important factor that contributes to the success of Web personalization in museums. It is a particularly important element, especially for audio guides, which must offer a certain level of flexibility in order to adapt the contents to the needs and interests of the users, just like a real museum guide would do. It also helps online, making the visitors feel comfortable and oriented in the virtual space, through virtual avatars for example. Studies indicate that the “social metaphor represented through the presence of personalized animated characters (similar to real life people) can reduce anxiety associated with the use of computers” (Bertoletti et al., 2001).

Personalization could also be a useful tool in the creation and development of online communities for museums (Beler et al., 2004). In fact, thanks to personalized applications such as alerts, thematic newsletters, customizable calendars and recommendation systems providing tailored content to people with specific interests, museums can identify homogeneous communities of users with the same concerns and needs. Once these different online communities have been identified, it is in the museum’s interest to foster them by developing tools and services that aid them in their functioning, especially by stimulating communication. This is when personalization can assist once again. In particular, online forums (Bowen et al., 2003) can benefit from the introduction of personalizing features such as notification of debates or issues that might be of interest to the user, information about other users with interests on specified topics (facilitating the networking between community users), personalized news generation based on personal interests, etc. These kinds of personalized services can increase the value of the underlying museum’s “e-community” beyond a social networking environment: “the website becomes an attractive permanent home base for the individual rather than a detached place to go online to socialize or network, thus strengthening the relation between the user and the institution” (Case et al., 2003).

By providing targeted information to users with different profiles and interests, personalized systems are much more likely to satisfy the visitor, who, as a consequence, is stimulated to come back and reuse the system or to encourage other people to try it as well. This is why personalization is also a fundamental marketing tool for the development of visitor fidelity, as well as new audiences.

Personalization and Learning

Besides helping museums to respond to their usability, marketing and accessibility needs, personalization has much potential when it comes to stimulating learning, as underlined by Brusilovsky (1994) who, early in the development of
the Web, pointed out how personalization techniques could be an important form of support in education. The reasons for this are varied. First of all, visitor studies seem to confirm that learning is encouraged when the information provided is described in terms that the visitor can understand. Using different terms and concepts, that take into consideration the level of knowledge, age, education of the user, etc., can therefore improve the overall didactic experience. This is precisely what happens with personalized applications where the information delivered to the visitors often changes according to whether they are a child, an adult, a neophyte or an expert.

Research also indicates that learning is facilitated when the information provided makes reference to visitors’ “previous knowledge”; that is to say, to what people already know or to concepts already encountered during navigation or exploration (Falk & Dierking, 1992). This suggests that museums should focus on how to activate visitors’ prior knowledge if possible. One of the means at their disposal is personalization, which could open new and effective means for long-term learning by providing adaptive descriptions of artifacts based on objects or concepts that the visitor has already visited or explored. This is, for example, the case in projects like ILEX, Hyperaudio, HIPS and the Marble Museum’s Virtual Guide — see Filippini-Fantoni (2003) for descriptions — that, through dynamically generated text, provide personalized information taking into consideration the user’s history. The description of the object being viewed or selected can make use of comparisons and contrasts to previously viewed objects or concepts. By providing such coherent and contextualized information, modeled on the user interaction with the exhibition space as well as with the system itself, such applications have enormous potential from the learning point of view.

Another mechanism that can be used to justify the use of personalization to stimulate learning is “subsequent experience” (Falk & Dierking, 1992). A number of researchers have hypothesized that repetition is the major mechanism for retaining memories over a long period of time (Brown & Kulick, 1997). This is why, by allowing the visitor to bookmark objects or concepts of interest during their navigation in the virtual or real environment and to explore them more in detail subsequently (see later for further information), personalization can make it possible to further deepen and continue the learning process from home by creating continuity between the visit and post-visit experiences.

Last but not least, learning is stimulated when a person can pursue their individual interests. Researchers distinguish between “situational interest” and “individual interest,” the first being defined as “the stimulus that occurs when one encounters tasks or environments with a certain degree of uncertainty, challenge or novelty” (Csikszentmihalyi & Hermanson, 1995). This is, for example, the case for museums where the presence of incentives like surprise, complexity and ambiguity lead to motivational states that result in curiosity and exploratory behavior (Csikszentmihalyi & Hermanson, 1995).

However this is not enough to guarantee that the visitor is actually stimulated to learn. In order for this to happen, museums have to attempt to respond to their visitors’ “individual interests,” that is “their preference for certain topics, subject areas or activities” (Hidi, 1990), as the pursuit of individual interests is usually associated with increased knowledge, positive emotions and the intrinsic desire to learn more. Personalizing an educational activity in terms of themes, objects or characters of high prior interest to students should therefore enhance the overall learning experience. Take, for example, those personalized applications (see later for details) that provide tailor-made visitor plans with consideration of the individual interests of a single visitor or a group of visitors. By suggesting artifacts relating to the visitor’s individual curiosity, the visit is more likely to result in fruitful learning activity.
In conclusion, by providing information at the right level of detail, stimulating subsequent experiences and taking into consideration individual interests as well as prior knowledge, personalization represents an excellent tool for all those educators wishing to stimulate and facilitate learning. This is why personalization techniques are often exploited in the creation of formal e-learning applications such as long-distance courses that are able to adapt to the student’s level of knowledge, cognitive preferences and interests, etc. For example, see the AHA Project on Adaptive Hypermedia for All [aha.win.tue.nl] at the Technical University of Eindhoven, The Netherlands, and the European IST ELENA Project on Enhanced Learning for Evolutive Neural Architectures [www.elena-project.org].

However, personalization can be also applied to more informal e-learning solutions like the ones that are often available on museums' Web sites or interactive devices, which, although not being actual lessons, represent very useful educational experiences that contribute to increasing the visitor’s knowledge and understanding about a specific issue.

WEB PERSONALIZATION FOR SCIENCE MUSEUMS

Until now, we have discussed more general issues concerning the use of personalization techniques in museums, focusing in particular on its potential to stimulate and facilitate the learning experience. In this section we consider some examples of how science museums in particular are applying these principles online. In fact, even if science museums are not the only cultural institutions to have experimented with personalization both online and on-site in the past few years — for a more general description of personalized applications in museums see Bowen and Filippini-Fantoni (2004) — they are among the ones that have expressed the strongest interest in these techniques. This is because science museums and science centers, whose exhibits are designed to promote playful exploration and discovery of scientific phenomena, have always been relatively aggressive adopters of information technology and innovative approaches; as a consequence, they have also been more eager to experiment with personalization.

Some museums have been focusing more on the usability and marketing aspects of personalization privileging applications such as personalized agendas, alerts and newsletters, which, although having an intrinsic pedagogical value, seem to focus more on promotion. However, science museums have been among the first to understand the real value of personalization as a learning tool, concentrating particularly on stimulating “subsequent experience,” “previous knowledge,” and “individual interest” in such a way as to explicitly encourage the continuity between the pre-visit, visit and post-visit experiences.

The first examples of Web personalization in a museum context were developed in the late 1990s in strict relation with the affirmation of academic research on adaptive hypermedia. Among them (Bowen & Filippini-Fantoni, 2004) was the SAGRES system (sagres.mct.pucrs.br), developed in 1999 by the Museum of Sciences and Technology of PUCRS (MCT), Porto Alegre, Brazil.

The SAGRES system (Bertoletti, 1999; Moraes, 1999) is an educational environment that presents the museum’s content adapted to the user’s characteristics (capacities and preferences). Based on information provided directly by the user or by the teacher (for students), the system determines the group of links appropriate to the user(s) and presents them in a personalized Web page.

The principle behind the project was an attempt to overcome the limitations implicit in the one fits all approach and to take the user’s individual interests as well as their level of knowledge into consideration when delivering information, with the aim of improving the overall learning expe-
Personalization Issues for Science Museum Web Sites and E-Learning

This is possible through an adaptation process that first generates a user model, based on information provided by the user. Once these data about the user have been collected, the adaptation process can select different types of documents conforming to the visitor’s model. This results in a dynamically generated HTML page with links pointing to personalized information: the page is created dynamically during the interaction of the user with the system and presents links to the documents, as well as connections to the communication mural (where users can interact with each other), to the document edition, and to the activities the user should perform (in the case of a group visit).

As well as being designed for individual users, the system is particularly meant for use in an educational setting. Through SAGRES, teachers are given the opportunity to define and register their students’ profiles, to accompany them and to evaluate their performance during the visit, using reports delivered by the system. At the same time, students are allowed to interchange ideas with colleagues in their groups and to work on the activities and subjects determined by the teacher.

PERSONALIZED VIRTUAL WEB SPACES

The main aim of the SAGRES project was to facilitate learning through the provision of information adapted to the level of knowledge and interest of the user. Since then, other methods have been adopted to guarantee a similar outcome.

Figure 1. The architecture of the SAGRES system
Various science museums, for example, provide users with tools that allow them to save images, articles, links, search results, forum discussion topics, as well as other types of information during navigation of the Web site. By doing so, the user creates a personal environment within the museum’s Web site, where they can return, find specific information of interest, and to which new items can be continuously added. This environment can be further equipped with other personalized services such as individual agendas or the ability to send personal e-cards.

Once the page has been created, visitors can log in every time they access the Web site to find all the information they need. By doing so, the user has the chance not only to find information of interest more easily, but also and especially to strengthen the learning process through reuse and repetition. The learning value of these applications for certain categories of users such as students and teachers is even greater. The personal space can offer teachers the possibility of suggesting of exhibits for their students to visit and questions that they would like the students to answer during the exploration. In response, the students can save links to the exhibits that most interest them, as well as making short notes both about questions they had at the beginning and about new questions that arise during the exploration.

One of the most interesting examples of this type of application is provided by the Ingenious project, undertaken by the National Museum of Science and Industry group in the United Kingdom and funded by the UK New Opportunities Fund (NOF) (www.nmsi.ac.uk/nmsipages/nofdigitise.asp). This project, online from mid-2004, aims at creating a learning environment for the public from the digitized collections of the Science Museum (London), the National Railway Museum (York) and the National Museum of Photography, Film and Television (Bradford) in the UK. Users of the Ingenious Web site (www.ingenious.org.uk) can explore and discover the rich collections of these museums through 50 narrative topics.

Figure 2. Ingenious home page
and over 30,000 images and other content-rich resources, such as library and object records. In addition visitors are provided with tools for entering a topical debate and personalizing their experience in the so called “CREATE” area, where registered users can save images and/or links from the debate areas, read sections and search queries. The users can also send personalized e-cards of images by e-mail and create a personal Web gallery from their bookmarked images, including the ability to incorporate personal comments that can be e-mailed to friends and colleagues.

Figure 2 shows a general shot of the Ingenious home page. The facilities include “My E-cards”

---

**Figure 3. Ingenious electronic cards**

---

**Figure 4. Ingenious selected hyperlinks**
Figure 5. Ingenious saved images

Figure 6. Ingenious Web galleries
to sent electronic cards (Figure 3), selected hyperlinks (Figure 4), saved images (Figure 5) and Web galleries (Figure 6).

Even if in the wider picture for Ingenious users, the umbrella group is lifelong learners, the application can be particularly suitable for older age school children, teachers, and researchers who could first explore a topic in the “read” or “see” sections of the site, then use the “save image” and e-card features and gradually progress to Web gallery tools for creating a personal resource. The Web gallery outcome would be used for a project, research, shared among a group of subject enthusiasts or a class (for instance). Community building could follow from this, through the usage of the debate features available on the site.

THE POST-VISIT EXPERIENCE

In some cases, personal virtual spaces can also include information about a visitor’s actual visit to museums, thus creating a direct link between the visit and the post-visit experience. Personalization is an effective tool for stimulating visitors at home to follow up on what caught their attention during the exhibition through a museum’s Web site. For example, the London Science Museum’s “In touch” project allows a record of a visitor’s interaction with various exhibits in the Wellcome Wing including an eye scan, voice, face and fingerprint recognition, photo editing, etc., to be recorded using their fingerprint as an identifier, thus avoiding the need for any physical ticket (www.sciencemuseumintouch.org.uk). The results are made available as part of a personal space within the museum’s Web site that can be accessed via the visitor’s first name and birth date.

Since 2000, when the project was originally implemented, Joe Cutting of the Science Museum reports that (as of January 2004) more than 400,000 Web pages have been created, of which around 8% have been accessed at least once. In order to simplify the system, reduce the operational problems that derive from such a large database, and increase the percentage of visitors using it, the museum has decided to replace the fingerprint method (which is not completely reliable in practice) by “an email it to me” option by the end of 2004. Every time a person wants to save one of the interactions, an e-mail address will have to be provided. By doing so, there will be no more automatically generated personal pages for the visitors. However, the museum is considering the inclusion of a link in the e-mail that would allow the visitors to set up a personal page if they wish. In this way only those who are really interested will set up a page and the museum will not have to maintain a huge and largely unused database. Figure 7 shows two screenshots from the exhibition itself and Figure 8 shows example pages from the associated Web site.

In a similar manner, the Visite Plus service offered by the Cité des Sciences et de l’Industrie (www.cite-sciences.fr) in Paris, which has been used on a number of successive temporary exhibitions, “Le Cerveau Intime,” “Le Canada Vraiment,” and “Opération Carbone,” allows the visitor to configure a personal profile (with information on preferred language, disabilities, etc.) on an interactive kiosk placed at the beginning of the exhibition through a special bar-coded ticket or on a PDA (Personal Digital Assistant). This data can then be used to access adapted information from the different interactive devices and to play various games and quizzes in the exhibitions. The results of such interaction, as well as the path followed by the visitor, are automatically saved by the Visite Plus system on a personal Web page, accessible on the museum’s Web site after the visit through the number of the bar-coded ticket or PDA. In this way, the visitors are able to analyze in more depth the subjects that particularly interested them during the exhibition (through the provision of additional information) and to compare results of their interactions with those of other visitors.
Figure 7. “In touch” exhibition screen shots

Figure 8. “In touch” Web pages
The fact that an important part of the content concerning the exhibition is accessible after the actual visit, at home or in another context, allows the visitor to focus more on experimentation and discovery while in the museum and to leave the more traditional didactic aspects for later. The Visite Plus system also offers the possibility of subscribing to a personalized periodical newsletter that focuses on a series of themes selected by the visitor at the moment of the registration. Options include selecting from a list of available subjects or receiving a complete dossier of the exhibition. See Figure 9 for an example of the view of the exhibit from the personalized Web site. Each square corresponds to a content area in the exhibition. The squares that are in full color represent the ones that have been accessed during the visit to the exhibition while the white ones correspond to the ones that have not been visited.

Similar concepts have been introduced and tested in the framework of the Electronic Guidebook Research Project (www.exploratorium.edu/guidebook), which began in 1998 at the San Francisco Exploratorium in California, in partnership with Hewlett-Packard Laboratories and the Concord Consortium. This is aimed at developing a roving resource to enhance a visitor’s experience at the museum (Hsi, 2003). In particular, the purpose of the project is to investigate how a mobile computing infrastructure enables museum visitors to create their own “guide” to the Exploratorium, using a personalized interactive system. This helps in better planning of their visit, getting the most out of it while they are in the museum and enabling reference back to it once they have returned to their home or classroom. The guidebook allows users to construct a record of their visit by bookmarking exhibit content,

Figure 9. Visite Plus personalized Web site
taking digital pictures from a camera near the exhibit, and accessing this information later on a personal “MyExploratorium” Web page in the museum or after their visit (Figure 10).

The project was designed as a proof of concept study to explore potential avenues for future research and development and therefore was not envisioned to support the implementation of a fully functional system. Nevertheless, the tests that have been run so far revealed interesting conclusions. Above all, the visitors liked the idea of being able to bookmark information for later reference. Both teachers and pupils thought this feature would allow the children to play more during their museum visit, completing related homework assignments after the visit (Semper & Spasojevic, 2002).

THE PRE-VISIT EXPERIENCE

The link between visit and post-visit experience can be also extended to the pre-visit phase through the implementation of systems that allow visitors to create personalized tours based on their interests and needs. Most museum visitors, even those who have not visited before, arrive with expectations about what will happen during the visit. Such hopes might concern specific subjects of interest that the person wants to explore, the physical characteristics of the museum, the types of activities that can be undertaken or the social context in which the exploration takes place (alone, as a family, within a larger organized group, etc.). All these factors merge to create a visitor’s personal agenda (Falk & Dierking, 1992). The success of the museum experience is partially
Personalization Issues for Science Museum Web Sites and E-Learning

defined by how well it corresponds to the visitor’s personal agenda.

Personalization is a useful tool to create such a correlation because it helps a visitor to find out what, within the museum, could fit better with their personal agenda or correspond more to their expectations. This can be done either from home on the museum’s Web site or directly on-site through interactive devices available in the museum. Upon completing a profile, where the intending visitor must indicate different types of information such as how and when they are tentatively planning to visit, with whom, how long they plan to stay, what sort of interest(s) they have and which language they understand, the system will be able to provide a personalized plan for the visit that takes the submitted information into consideration. Personalized museum plans can be very useful, especially for large museums where visitors are likely to be overwhelmed by the number of objects or exhibits available for viewing during a single visit. In such a context, visitors are often disoriented and find it difficult to decide what they want to see or do. Answering a few very simple questions, or defining a few criteria, can help them to overcome these limitations, enjoy the visit more fully and learn more easily.

A number of museums are working on developing online and onsite applications based on these principles. The National Museum of Ethnology in Leiden, the Netherlands (www.rmv.nl), for example, has developed an onsite facility called “The tour of the world in 80 questions” that allows children aged 7 to 13 to print out a personalized tour plan of the museum based on an individual choice of subjects and continents. The tour plan, which is colorful and easy to understand for children, includes a series of maps that help locate the objects, a brief description of the artifacts and a list of questions related to the subjects chosen, which the young visitors need to answer during their museum exploration.

The Cité des Sciences et de L’Industrie in Paris is undertaking a project called “Navigateur” (Navigator), which will allow visitors to create a personalized tour based on an individual choice amid a set of criteria which include the context of the visit (alone, family, group), the language spoken, the particular interests, the time available and the type of experience desired. Once the visitor has set the criteria that are most relevant for them and has checked the offerings on the museum interactive plan, the personalized proposal can be saved on the museum bar-coded ticket, which will be used during the actual visit, when using different interactive devices throughout the museum, to obtain further assistance in finding the recommended exhibits or to reset the criteria based on new interests that might have arisen during exploration. The system will be linked directly with Visit Plus, thus creating continuity between the pre-visit, actual visit and post-visit experience, through the use of personalization.

CONCLUSION

The examples provided here from different science museums all over the world help to prove the potential role that personalization could play in strengthening the overall learning process before, during and after the actual visit, in advance through activities that orient visitors and afterwards through opportunities to continue reflection and explore related ideas. However, despite the obvious potential benefits that these applications can bring to the visitor’s experiences, there is still very little evidence that these systems work in the terms envisaged by their promoters, especially with respect to learning. This is because, due to their relatively recent nature, most of these projects have not yet been subjected to thorough evaluations that focus on establishing, among other things, the long-lasting effects of personalization on the learning process. Until now,
Personalization Issues for Science Museum Web Sites and E-Learning

the very few evaluations that have been carried out have focused mainly on whether people use the systems or not, why they do so, where they encounter most difficulties and on their usability in general. Despite the fact that further studies are needed in order to shed light on the effectiveness of personalization as a pedagogical tool, the first evaluations of these early examples, as well as other similar projects, have given initial help in indicating various pros and cons related to their use.

The overall feedback concerning the introduction of personalizing applications to audio guides and virtual environments seems to be reasonably positive: visitors are spending more time in the virtual and real museum, they access information at the level of detail desired and appreciate the idea of being able to bookmark information for reference later (Semper & Spasojevic, 2002). In particular, a study by Cordoba and Lepper (1996) has evaluated the consequences of personalization with respect to stimulating intrinsic motivation and learning in a computer-based educational environment. The findings provide strong evidence that the students for whom the learning contexts had been personalized, through the incorporation of incidental individualized information about their backgrounds and interests, displayed better gains in motivation, involvement and learning than their counterparts for whom the contexts had not been personalized.

However some drawbacks have also emerged. First of all, there are the issues related to the difficulty and expense of implementation and also problems in practical use by visitors. So far it seems that only a limited number of visitors take advantage of the benefits available through personalization, partly because the systems are not implemented in a clear and easy manner and partly because most visitors are either not ready for technology or not willing to invest time in it. Therefore it is important to remember that personalization should not be implemented for the sake of it but when and because it brings added value to the museum for, if not all, a good percentage of visitors. Only if this occurs can the costs for investment and development be justified.

Some experts have warned against the use of personalization. Nielsen (1998) has argued that personalization is over-rated, saying that good basic Web navigation is much more important. For example, it is helpful to consider different classes of use in the main home page, such as physical visitors, the disabled, children, teachers, researchers, groups, etc., and to give each of these a relevant view of the resources that are available (Bowen & Bowen, 2000). Such usability issues are certainly important, and relatively cheap to address with good design, but even Nielsen admits that there are special cases were personalization is useful.

More recently, there have been further questions about the effectiveness of personalization (Festa, 2003; McGovern, 2003), despite the enthusiasm of some. For example, the costs may be up to four times that of a normal Web site, around a quarter of users may actually avoid personalized Web sites due to privacy concerns and only 8% are encouraged to revisit because of personalized facilities (Jupiter Research, 2003). This compares with 54% who considered fast-loading pages and 52% who rate better navigation as being important. However, other surveys indicate that personalization can be effective, for example in the field of downloadable music (Tam & Ho, 2003).

Another issue that needs to be stressed in personalization is related to standardization procedures and applications. This process is central both for content description and user profile definition using metadata (Conlan et al., 2002). The description process can however be very time-consuming and expensive, but if it is pursued properly it allows the resources to be reused for different purposes and a visitor profile to be created using various different sources of information following evaluation criteria. Museums are sometimes not very quick in adopting new technologies but in some cases the slow perspective allows them to
make the most of other institutions’ initial mistakes and thus to avoid them. Involvement with standards provides a good opportunity to share such knowledge.

Thus it is recommended for museums to use personalization on Web sites judiciously at the moment, although science museums with good funding may wish to be more adventurous. There is a place for personalization in leading-edge Web sites and for certain innovative facilities like advanced Web support for specific exhibits. It is an area that museums should certainly consider, but the costs should be weighed against the benefits. Of course, the costs are likely to decrease as commercial and open source support improves in this area. At the moment, not insignificant development effort is needed for such facilities, but in the future they could be increasingly packaged with standard database-oriented Web support software, such as content management systems, as understanding of what is useful and not useful is gained from practical experience. This is certainly an interesting and fast-moving area that should be monitored by innovative science museums, especially at a national level.

REFERENCES


Personalization Issues for Science Museum Web Sites and E-Learning


ENDNOTES

1 A “cookie” is a small piece of data sent by a website and stored on the client-side (browser) computer that can be reused later on the server-side (the Web site that sent the cookie) as unique information concerning a user.

2 A Web server log is a record of each access to a Web server with information such as the name of the client computer, the date/time and the resource accessed.

3 These applications are currently available on a number of different museums’ Web sites such as the Metropolitan Museum of Art, the Whitney Museum of American Art, etc. For a detailed description of these applications (see Bowen & Filippini-Fantoni, 2004).

4 Please note that the distinction between formal and informal education is used here in a rather loose sense. Usually, in the educational sector, classrooms are considered formal learning settings, while museums are considered informal learning settings. As an alternative, we propose here to use the term formal e-learning tools in relation to proper courses meant for students who cannot attend classes; while by informal e-learning tools we refer to online or onsite educational environments.

5 Note that the acquisition of knowledge about the visitor is done in an explicit way: information is directly extracted, through the filling of forms, with direct answers to questionnaires. SAGRES works with two kinds of models: individual model and group model. The group model is built by the teacher and used by students. The teacher
is responsible for the definition of the students’ characteristics, by the definition of the group stereotype (subject, knowledge level and language of the consultation), the activities stereotypes and the classes (name of the students presented in the group).

It is not the intention of this chapter to be negative towards the use of personalization techniques in museums, but to highlight constructively some of the questions that come to light when the social uses and design problems are considered.

For more detailed information on the problems related to the implementation and use of personalization techniques see Filippini-Fantoni (2003).
Chapter 4.15
Personalization and Customer Satisfaction in Mobile Commerce

HY Sonya Hsu
Southern Illinois University, USA

Songpol Kulviwat
Hofstra University, USA

INTRODUCTION

The advancement of wireless technology facilitates both consumers’ activities and business transactions. With the rapid proliferation and widespread use of mobile devices, including mobile phones, personal digital assistants (PDAs), and handheld computers, mobile commerce or m-commerce is widely considered to be a driving force for the next generation of electronic commerce (e-commerce). According to Jupiter Research, the m-commerce industry is expected to be US$22 billion globally by 2005. However, to date many promising technologies—especially m-commerce applications—have failed with the notable exceptions of i-Mode service and short messaging service (SMS).

Popular “i-Mode”, produced by NTT DoCoMo of Japan, is a service that enables wireless Web browsing and e-mail from mobile phones. The “i-Mode service” has been the first successful commercial introduction of 3G (third-generation) mobile applications. It exceeded expectations and acquired over 30 million profitable users in a three-year period (Cohen, 2002).

One of the main goals of most operators might be building customer satisfaction and loyalty by providing one or more ‘killer apps’ to them. One way is to integrate customer relationship management (CRM) into the development of mobile services’ applications. Some firms have tried to target these applications to their customers on an individualized basis. “Personalization” may be the way to achieve that. Specifically, personalization can be regarded as the use of technology and user/customer information to match multimedia content with individual needs with the goal of producing user satisfaction. Personalization can
be presented by an IP services framework that allows operators and subscribers through self-service provisioning approaches to control the types of service and applications they want and are willing to buy.

The purpose of this article is to develop a deeper understanding of personalization, with an emphasis on those factors that lead to customer satisfaction and/or delight. Specifically, this article presents factors contributing to consequences derived from using personalized applications and services in m-commerce.

BACKGROUND

In their pilot study, Ho and Kwok (2003) applied the technology acceptance model (TAM) originated by Davis (1989) to their m-commerce study. They utilized four constructs to predict the service subscribers’ intention to switch: number of generalized messages, perceived ease of use of general advertisements, perceived usefulness of personalized message, and privacy issues about personalized advertisements.

This article extends the thrust of Ho and Kwok’s research to incorporate the effect of personalization on customers’ satisfaction and delight that could contribute to CRM. Customers’ satisfaction and delight are derived from expectancy theory, and they are discussed by Oliver (1981), Oliver, Rust, and Varki (1997), Spreng, Mackenzie, and Olshavsky (1996), and Verma (2003).

Expectancy: Satisfaction and Delight

Expectancy theory is used to frame the evaluation of mobile services users. Oliver (1981) defined expectation to include two components: the probability of occurrence (e.g., the likelihood that a personalized cell service will be available) and an evaluation of the occurrence (e.g., the degree to which the personalization level is desirable or undesirable). The disconfirmation/confirmation paradigm of satisfaction is based on expectancy theory. It can be an emotional response to the comparison of the performance received and the products’ normative standards. When the performance and expectations are at variance with each other, there is a discrepancy. This discrepancy could be either positive (when performance exceeds the expectations), which often causes satisfied state, or it could be negative, when performance is worse off than expected (Oliver, 1981). In other words, the consumer would be satisfied if perceptions match expectations or if confirmations are reached. Consistent with Spreng et al. (1996), satisfaction arises when consumers compare their perceptions of the performance of a good and/or service to both their desires and expectations. As such, satisfaction is a subjective judgment and may imply mere fulfillment.

Delight is a positively valence state reflecting high levels of consumption-based affect. The feeling of delight is experienced when the customer is pleasantly surprised in response to an experienced disconfirmation. It is the feeling state containing high levels of joy and surprise (Westbrook & Oliver, 1991). Further, Oliver et al. (1997) proposed and confirmed that delight is a function of surprising consumption, arousal, and positive effect or a function of surprisingly unexpected pleasure. They empirically confirmed that delight is a “mixture” of positive effect and arousal or surprise. It is associated with the level of arousal intensity. Moreover, it is a reaction experienced by the customer when he or she receives a service and/or a good that does not simply evoke a feeling of satisfaction, but also provides an unexpected value or unanticipated additional pleasure. In other words, delight occurs when the outcome is unanticipated or surprising. It can be marked by pleasurable, unforgettable, and memorable feelings in a service encounter or a product purchase (Verma, 2003). It is thought to be the key to customer loyalty and loyalty-driven profit (Oliver et al., 1997) and is known as the highest level of expectation-disconfirmation paradigm.
Technology Acceptance Model (TAM)

From Davis’ (1989) TAM model, ease of use (EOU), and perceived usefulness (PU) of a technology are factors that either directly or indirectly increase a person’s intention to adopt an innovation. While perceived usefulness is the degree to which a person believes that using a particular technology/system would enhance the outcome performance, perceived ease of use is the extent to which a person believes that using a particular technology/system will be free of effort (Davis, 1989). TAM could be helpful in predicting the usage of personalized applications and services. Greer and Murtaza (2003) adapted the TAM model to study issues that impact the valuation of Web personalization as well as factors that determine customer use of Web personalization. Ho and Kwok (2003) adapted Davis’ (1989) EOU and supported the effect of using a generalized message on changing a service provider. They also used “PU of personalized service” to test the importance of personalization in mobile commerce. They found support for both. Most importantly, the PU of personalized service was the most effective factor, together with ease of locating generalized message and the amount of generalized message that affected the decision to change to a new service (Ho & Kwok, 2003).

MAIN THRUST OF THE ARTICLE

Usually when there are too many generalized messages, customers lose their motivation to read, retrieve, or even locate a useful message. In addition, the amount of space available on the mobile screen limits the amount of options and information. Given this, personalization is considered to be the key factor for success/failure of mobile devices and services. Information and services must become increasingly tailored to individual user preferences and characteristics in order to accommodate limited space and scarce airtime. Personalization is viewed as including “recognition of a customer’s uniqueness” (Surprenant & Solomon, 1987, p. 87), use of a customer’s name, and response to customer needs (Goodwin & Smith, 1990).

Message Format

Carlson et al. (1998) characterized medium richness as the capacity to convey information. It is further defined as the ability to provide immediate feedback to customers’ consumption of media. Rich information can be produced by giving immediate feedback, having a variety of available communication cues, understandable/common language, and foremost, personalization of the medium (Carlson et al., 1998).

Media richness theory postulates that media selection depends on the uncertainty of the task at hand (Kumar & Benbasat, 2002). Both media richness theory and the TAM model have illustrated their relationships with task orientation. Also, social presence theory postulates a particular communication task based on the degree of necessary social presence that links a selection of media (Kumar & Benbasat, 2002). Originally, it referred to the degree to which a medium allows a user to establish a personal connection with the other users. Social presence seems to be moving towards a task orientation at an individual level in the latter theoretical development, such as the para-social concept from Kumar and Benbasat (2002). Para-social is a combination product of social presence and media richness. This article focuses on the PU of personalized messages that employ a task orientation, while two different formats of messages (text and multimedia) were drawn from media richness theory.

Personalization

Personalization can be defined as the use of technology and user/customer information to customize multimedia content so as to match with individual needs and ultimately produce
Personalization translates individual profiles into unique presentations. The individual profiles can be built upon user preferences, the quality of his or her senses, user location/environment, contexts, users’ network, and terminal capabilities. Morris-Lee’s (2002) study on personalization of brochures indicated that personalization helped increase interest and involvement. The more personalized features are, the greater the possibility of increased costs (Greer & Murtaza, 2003). Hence, these increasing costs hopefully should produce greater customer satisfaction and retention, thus a greater return. This is a very important point because, for example, a 5% increase in customer retention costs can translate into a 25%-125% increase in company profitability (Reichheld et al., 2001). Also, personalization of service has been found to have a positive impact on customers’ evaluations of service encounters (Surprenant & Solomon, 1987).

Figure 1 presents the research model of personalization in m-commerce. The model is developed based on the TAM (Davis, 1989) and expectancy theory (Oliver, 1981). Specifically, this research model extends Ho and Kwok’s (2003) research framework. Further, our model integrates customer expectancy as an endogenous variable. The model includes six sets of variables: (1) number of generalized messages, (2) perceived usefulness, (3) text messages, (4) multimedia message, (5) satisfaction, and (6) delight.

**FUTURE TRENDS**

As predicted, text message predictor had a positive association with the dependent variable. Number of generalized messages was negatively related to satisfaction. Multimedia message was not a significant predictor of satisfaction and was deleted in the second model. On the other hand, the analyzes of “delight,” “multimedia message” contributed the most importance to the
equations and was positively related to delight. Text message and PU of personalized message had positive associations with delight, while the number of generalized messages had a negative contribution to the equation (Hsu, Bruner, & Kulviwat, 2005).

According to Santos et al. (2003), even though satisfaction and delight are two different constructs, each serves a dimension of confirmation (satisfaction) at one end and disconfirmation (delight) on the other end. Increasing literature has being drawn in the difference between consumer satisfaction and delight (Kumar & Olshavsky, 1997; Oliver et al., 1997). To compare satisfaction with delight, Oliver et al. (1997) see customer delight as being fundamentally different from customer satisfaction. Compared to satisfaction, delight seems more abstract and more extreme in terms of affection. While satisfaction may be induced by avoiding problems or may meet standard/minimum requirement, delight requires more than that. Oliver et al. (1997) empirically confirmed the distinction between the satisfaction and delight constructs, with delight being a higher level of satisfaction. In fact, customer delight is associated with a strong and positive emotional reaction to a product or service. Thus, both practitioners and scholars should manage customer delight as a separate goal from satisfaction.

Te’eni, Sagie, Schwartz, Zaidman, and Amichai-Hamburger (2001) used three dimensions to define media richness further; these are interactivity, adaptiveness, and channel capacity. Beyond just a different format from a text message, future researchers may look into a deeper understanding of multimedia messages that convey information for possible customers’ delight in addition to satisfaction. Delightedness can be marked with pleasurable, unforgettable, and memorable where customer loyalty is rooted (Verma, 2003; Oliver et al., 1997). With the limitation of student population, future research may investigate some professional group that has reasons to use mobile commerce and/or some population that has more disposable income at hand. Another limitation that can also be addressed in the future is the sophistication of multimedia services and the maturity of users. In other words, future researchers may look into some markets that have rolled out the mMode of AT&T, Mobile Web of Verizon, and/or Sprint’s PCS.

From mobile application point of view, “personalization” can be more sensitive to users’ needs, for example, location-based application as in www.mobull.usf.edu. Local merchants ally to deliver a personalized text message—such as sales, promotion advertisement, coupons—from a Web site to a wireless device based on personal preferences that are set up by each individual. Location-based services utilize location information to provide specialized contents to mobile users (Varshney, 2003). Explicit user permissions should be obtained before “pushing” any advertising contents to particular users (Varshney, 2003). Push and pull advertisement, of course, relates to the issues of privacy and sharing of user information. Therefore, the “trust” matter may surface between a group of local merchants and individual consumers.

**CONCLUSION**

This article identifies the same situation as in Ho and Kwok (2003) that the amount of generalized message had a negative effect on customer satisfaction. Personalized message is more likely related to customer satisfaction and delight. The TAM model and expectancy theory were drawn as the foundation of this research model. Media richness explains the division between text message and multimedia message, whereas TAM contributes the perceived usefulness of personalized message.

Beyond personalization, this article attempts to merge the media richness theory with expectancy
Personalization and Customer Satisfaction in Mobile Commerce

theory. Specifically, it explains the relationships between text/multimedia message and customer’s satisfaction/delight. The article concludes that consumers would like to have a richer media to experience a “delightful” emotion. Consistent with the principle of media richness (Carlson et al., 1998): the more complex media format, the more information can be delivered in a message. If managers would like to increase effectiveness and/or efficiency of mobile services, text message alone would not be sufficient for market differentiation to gain competitive advantage. With personalization, multimedia formats can be a supplement tool to increase the interaction with consumers when launching advertising campaigns. The richer the media, the more effective it is in communication.

REFERENCES


**KEY TERMS**

**Customer Delight:** The feeling of delight is experienced when the customer is pleasantly surprised in response to an experienced disconfirmation.

**Customer Satisfaction:** Based on the consumption, consumer would be satisfied if perceptions match expectations or if confirmations are reached.

**Expectance Theory:** Oliver defined expectation to include two components: the probability of occurrence and an evaluation of the occurrence. The discrepancy of confirmation could be either positive or negative.

**Media Richness:** Media richness theory postulates that media selection depends on the uncertainty of the task at hand. The more complex media format, the more information can be delivered in a message.

**Mobile or M-Commerce:** Both consumers’ activities and business transactions are facilitated by the advancement of wireless technology including cellular phones, wireless PDAs, or any hand-held units.

**Personalization:** Can be regarded as services of the use of technology and user/customer information to customize multimedia content aiming to match with individual needs and ultimately deliver customers’ or users’ satisfaction.

**Technology Acceptance Model (TAM):** From Davis’ TAM model, ease of use (EOU) and perceived usefulness (PU) of a technology are factors that either directly or indirectly increase a person’s intention to adopt an innovation.
Chapter 4.16
Integrating Human Computer Interaction in Veterinary Medicine Curricula

Gale Parchoma
University of Saskatchewan, Canada

Susan M. Taylor
University of Saskatchewan, Canada

Jonathan M. Naylor
Ross University School of Veterinary Medicine, West Indies

Sameeh M. Abutarbush
University of Prince Edward Island, Canada

Katharina L. Lohmann
University of Saskatchewan, Canada

Kathy Schwarz
University of Saskatchewan, Canada

Cheryl Waldner
University of Saskatchewan, Canada

Sharon Porterfield
University of Saskatchewan, Canada

Cindy L. Shmon
University of Saskatchewan, Canada

Lydden Polley
University of Saskatchewan, Canada

Chris Clark
University of Saskatchewan, Canada

ABSTRACT

This chapter discusses contemporary global challenges facing veterinary educators and summarizes some of the economic, social, political, and technological pressures underlying curricular and pedagogical change initiatives. Integrating human computer interaction (HCI) into veterinary medicine curricula, as a strategy for implementing pedagogical transformation, is reviewed. Computer-assisted learning (CAL) projects recently developed at a veterinary college are described. Results of studies evaluating the effectiveness of CAL approaches to HCI integration within the veterinary medicine curricula are reported. Future research directions are proposed.
**INTRODUCTION**

Contemporary veterinary medical education is in a transformative state. Veterinary educators are responding to public demands for the expansion of veterinary roles, for specialized veterinary care, and increased concern for animal welfare, global demands for standardization of veterinary curricula, and veterinary teaching hospital practice in the context of coping with a diminishing pool of academic veterinary researcher/educators who must manage a rapidly expanding knowledge base. Increasingly, veterinary educators are seeking human computer interaction (HCI) solutions to addressing these emergent challenges. This chapter examines these emergent challenges and describes international initiatives focused on integrating HCI into veterinary medical curricula. The chapter includes an in-depth examination of technology-enhanced learning (TEL) research and development program at the Western College of Veterinary Medicine in Canada.

**EMERGENT CHALLENGES**

Contemporary veterinary educators are responding to a series of emergent challenges. Increasing public demands to expand veterinary roles into public health-assurance issues are at the fore of these challenges. Veterinary responsibilities for ensuring a secure, sustainable food supply and managing industrial-scale food animal production—within a climate of public fear of pandemic disease outbreaks—have globalized these issues within veterinary colleges, regions, and governance bodies. Animal welfare concerns, as well as ecological and environmental, issues are affecting not only what is taught in veterinary school, but also the way in which it is taught. Public demand for access to specialized veterinary care and expanding pressures on veterinary teaching hospitals to train specialists and increase caseload have resulted in a predominance of secondary and tertiary cases (referrals from practicing veterinarians) being evaluated at veterinary teaching clinics, decreasing the access of veterinary students to “general practice” cases.

Meanwhile, the pool of veterinary educators is diminishing as more financially rewarding opportunities abound in private practice and the corporate sector. All veterinary colleges cannot secure, in a timely fashion, candidates for open positions in veterinary teaching. Therefore, student access to inter-institutional experts and specialists is an emerging necessity. The veterinary knowledge base is rapidly expanding, so that it is no longer realistic to teach veterinary students “everything they need to know” within their four-year curriculum. Debates about the merits of traditional broad-based curricula versus early specialization, the appropriateness of national and regional versus global credential standards, and even pedagogical approaches to veterinary teaching and learning are recurrent themes in current veterinary literature. The combined effects of these challenges for change in veterinary teaching methodologies have resulted in calls for HCI alternatives to traditional lecture-based pedagogy and to invasive animal use in veterinary educational laboratory exercises (Association of American Veterinary Medical Colleges, 2005; Fernandes, 2004).

**Globalization of the Veterinary Profession**

The veterinary medical profession is increasingly expected to contribute to the development of solutions for global problems. This new responsibility makes the teaching and practice of veterinary medicine a global concern. Societal expectations that the veterinary profession should “undertake roles relevant to the re-assurance of human well-being, in terms of public health,” and address “the increasing consciousness in issues of animal welfare, sustainable animal production and environmental protection” (Rodriguez-Martinez, 2004,
Integrating Human Computer Interaction in Veterinary Medicine Curricula

p. 30) are driving this change. Animal disease crises with implications for human health, such as Bovine Spongiform Encephalopathy (BSE; mad cow disease), Avian influenza, West Nile virus, and Chronic Wasting Disease have brought issues of “food safety, public health, food animal health, and food animal production” (Walsh, 2004, p. 9) to the fore of public attention. As food security, environmental sustainability, and biodiversity (Edwards, 2004) are not subject to national or regional boundaries, veterinary education curricula and credentialing standards have become subjects of international interest.

Concerns about global educational standards in the health science professions in general (Clarkson, 2005; Jackson & Callman, 2006), as well as in veterinary medical education in particular (Edwards, 2004; Hammick, 2005a) have influenced the adoption of evidence-based evaluation of teaching and learning effectiveness. Initiatives, such as the formation of the Best Evidence Medical Education (BEME) and Campbell collaborations, focus on systematic reviews of medical science education for the purpose of “providing teachers, institutions and all concerned with healthcare education” with evidence to support curricular and pedagogical decision-making (Hammick, 2005b, p. 339). Computerized collection, analysis, and dissemination of evidence supporting educational effectiveness have come to the fore in these efforts.

Veterinary Teaching Hospital Challenges

The clinical veterinary profession has dramatically changed during the last 20 years, with increasing public demands and willingness to pay for a higher standard of care and veterinary expertise. The role of specialists has increased, not only within particular species, such as small animals, horses, food animals, exotic animals, and so forth, but also within the disciplines, such as epidemiology, internal medicine, surgery, cardiology, dermatology, dentistry, ophthalmology, neurology, and oncology. Most clinicians teaching at veterinary teaching hospitals are board-certified specialists. Despite the need for veterinary students to “handle routine cases typical of those seen in veterinary practice,” in most teaching hospitals, “the majority of the caseload is based on referrals, and most cases could be classified as receiving secondary or tertiary care” (Brown, 2003, p. 227).

There are many different diseases affecting wildlife, exotic, companion, and food animal health. It is not possible to discuss all of the conditions and their possible clinical presentations within the current veterinary curriculum. Ensuring that all interested students see examples of each condition during their clinical rotations is also not feasible. As a result, learning opportunities are often “hit or miss” with respect to specific case scenarios to which the students are exposed. Computer-assisted learning resources provide one method of addressing this problem by maximizing the use of current and historical case materials to expose students to the “typical clinical presentations” of many common or important diseases in each species (Abutarbush, Naylor, Parchoma, D’Eon, Petrie, & Carruthers, 2006; Clark, 2005).

Information Expansion

Veterinary information has rapidly expanded (Naylor, 2005), making it impossible for veterinary educational programs to teach students everything they will need to know for their career during a four-year veterinary program (Rodriguez-Martinez, 2004). In addition, current information is constantly being updated and changing. As a result, veterinary curricula and the expected competencies of veterinary graduates require constant review (Edwards, 2004). It is a challenge to balance the demands for a broad based curriculum to support professional licensure and provide the “basic training that should be broad enough to warrant
Integrating Human Computer Interaction in Veterinary Medicine Curricula

the graduate to practice, in principle, in any field of veterinary medicine” while also satisfying the increasing demands for specialized knowledge and expertise within the limited timeframe of undergraduate study (Rodriguez-Martinez, 2004, p. 31; see also Fernandes, 2004). Furthermore, preparing students for clinical practice is only part of the picture. Veterinary curricula are also expected to prepare graduates for potential careers in epidemiology, pharmacology, public health, or basic science research. The futility of trying to “teach it all” is clear (Rodriguez-Martinez, 2004, p. 31). Veterinary curricula must make choices, and, above all, veterinary graduates must have the skills, attitudes, and dedication that promote life-long learning and continuing education after graduation.

Veterinary Educator Shortages

While the need for access to veterinary learning opportunities is expanding, the pool of veterinary educators is diminishing. Veterinary colleges “worldwide are facing numerous challenges including increasingly limited resource allocations, difficulties on enrollments, gender changes, keeping up with advances in information and other technologies, remaining aware and responsive to … [learners] and the need to aggressively globalize their teaching, research and outreach programs” (Fernandes, 2004, pp. 9-10). Difficult economic situations in developing countries further exacerbate efforts to meet these challenges. In developed countries, attracting potential faculty clinicians away from fiscally rewarding private specialty practices and positions in the veterinary corporate sector (e.g., food sector, feed, and pharmaceutical industries) into an under-resourced and increasingly demanding academic work environment is a pervasive concern (De Castrol & Zuconi, 2004). “[A]mong veterinary establishments worldwide [there] is [a] worrying lowering number of veterinary graduates enrolling in postgraduate veterinary training” (Rodriguez-Martinez, 2004, p. 36). Most veterinary colleges are actively struggling to replace their retiring professoriate as well as their young faculty who are being lured into non-academic careers.

The Pedagogical Shift

Undergraduate veterinary education in Europe (and perhaps worldwide) is probably still considered one of the most traditional university educations, with extensive formal teaching, based upon rigid curricula, most often teaching facts rather than principles (Rodriguez-Martinez, 2004, p. 32).

Teaching veterinary students the processes of independent and collaborative “problem-identification, problem-solving and decision-making” and transforming pedagogical praxis from traditional fact-based lectures and demonstrations into “active, student-centred learning environments” has become a priority in many veterinary curricula (Rodriguez-Martinez, 2004, pp. 31-32). Computer-assisted learning (CAL) plays an important role in this change initiative. CAL has the capacity to link dispersed veterinary specialists in knowledge-sharing and collaborative teaching ventures, and to provide students with independent and collaborative case analyses and problem solving opportunities.

The use of CAL in veterinary education is expanding, and a range of tools supporting problem-based learning in virtual environments has been developed (Dhein, 2005; Hines, Collin, Quitadamo, Brahler, Knudson, & Crouch, 2005; Schoenfeld-Tacher, Bright, McConnell, Marley, & Kogan, 2005). Based in the United Kingdom, the Computer-aided Learning in Veterinary Education (CLIVE) consortium involves 14 colleges of veterinary medicine around the globe dedicated to developing and sharing electronic resources for CAL. A range of veterinary CAL initiatives have also developed in North America. Examples include Cornell University’s Consultant, a diag-
nositic database “designed to link over 500 clinical signs and symptoms to nearly 7,000 possible diagnoses or disease conditions (White, 2005); the integration of HCI into pharmacology instruction (Kochevar, 2003); large animal veterinary education (Dascanio, 2003), antimicrobial resistance and animal welfare (Bernardo & Malinowski, 2005); systemic pathology (Hines et al., 2005), and an introduction to anesthetic and surgical principles and techniques (Howe, Boothe, & Hartsfield, 2005).

The Animal Behaviour Learning Environment (ABLE), developed in Australia, uses virtual case studies and Web-based resources to support veterinary students in their efforts to diagnose and prepare treatment plans (McGreevy, Della Torre, & Evans, 2003). The Virtual Veterinary Medicine Learning Commons, a Canadian collaboration among the Atlantic, Ontario, and Western Colleges of Veterinary Medicine, has resulted in the sharing of high quality case-based clinical instruction by advanced broadband Internet technology (V2VLC, 2001). The University of Montreal has produced multilingual bovine medicine CD-ROMs, (Carriére, DesCôteaux, Durocher, & Harvey, 2005; Desrochers & Harvey, 2002).

INTEGRATING HCI INTO VETERINARY MEDICAL EDUCATION AT THE UNIVERSITY OF SASKATCHEWAN

The Western College of Veterinary Medicine (WCVM), established in 1964 at the University of Saskatchewan, provides undergraduate and graduate veterinary education to students from the four western Canadian provinces, as well as to national and international students. WCVM is accredited by the Council on Education of the American Veterinary Medical Association. The stated aim of the Doctor of Veterinary Medicine Program at the Western College of Veterinary Medicine is “to provide an excellent education in animal health that meets society’s needs, now and in the future” (WCVM, 2006, Mission statement, para. 1). The WCVM undergraduate veterinary program currently provides a fairly traditional education. In the first three years, a strong foundation in the basic sciences is followed by clinical courses, while the final year consists of clinical rotations through various clinical specialty areas. Within the next three years, however, the WCVM will be moving to a new core-elective curriculum whereby the core content (required to be taken by every student) will be decreased in order to expand elective opportunities available to students. This will allow each student to modify the curriculum to meet his or her own needs. The broad-based core curriculum will encompass the common knowledge, skills, and behaviors that veterinary students need to acquire, regardless of their ultimate career objectives. This approach will provide students with a basic understanding of the common problems facing veterinarians and prepare them for the national licensing examination. Elective opportunities will complement and enhance the knowledge, skills, and behaviors acquired in the core, allowing students to gain the additional knowledge, ability, and experience necessary for confident and successful entry in their chosen career path.

Computer-assisted learning has been a component of the veterinary curriculum at the WCVM for many years, and it is anticipated that it will play an even more important role in the new curriculum. Some courses (Virology and Clinical Pharmacology) have used WebCT to provide students with electronic lecture handouts, copies of PowerPoint slide lectures, quizzes, and interactions with faculty. Other courses, including Small Animal Medicine and Surgery, have made case materials (photographs and videos) available to students prior to case-based or modified problem-based class discussions. A few case-based CD-ROMs and DVDs have also been developed to supplement lecture-based discussions. In the Department of Small Animal Clinical Sciences
two clinical faculty members (one surgeon and one internist) spent a sabbatical leave at CLIVE developing two interactive DVDs for teaching clinical neurology to veterinary students (Shmon & Taylor, 2003; Taylor & Shmon, 2003). In the Department of Large Animal Clinical Sciences, a series of CAL initiatives began in late 1990s. Results included interactive case simulations using the problem-oriented format: Diseases of Calves and Diseases of Horses (Naylor, 1996) and more recently CD-ROM-based learning modules concentrating on the skills of auscultation, The Art of Equine Auscultation, 2nd edition, (Naylor, 2001) and The Art of Bovine Auscultation (Naylor, 2003), and bovine foot care: Cattle Claw Care (Clark, 2004).

**Provincially Supported Technology-Enhanced Learning Initiatives in Veterinary Curricular Development and Educational Research**

Funding support from the province of Saskatchewan’s Technology Enhanced Learning (TEL) initiative from 2001 to the present has extended the scope of CAL development and evaluation of CAL projects undertaken at the WCVM.

In the Department of Veterinary Microbiology, a Web-based learning resource in veterinary parasitology is being developed. The key element of this resource is a searchable database containing information on the taxonomy, morphology, geographic and host distributions, life cycle, pathology, clinical signs, diagnosis, epidemiology, treatment, control, and public health significance of approximately 100 parasites important in domestic animals in Canada. The text material for each parasite is complemented by a set of images and by animations summarizing the life cycles. For veterinary students, probably the primary users, the database can be accessed through WebCT, which provides opportunities for discussions and quizzes, together with the potential to use the database as an integral component of the students’ learning experiences in parasitology. Practicing veterinarians, animal health technicians, and students in other veterinary programs will also be able to access the database directly through the Web, although it is password protected. To assess whether this WCVM parasitology resource has wider application in veterinary education, parasitologists at the School of Veterinary and Biomedical Sciences at Murdoch University in Western Australia are participating in its development and evaluation. Information and expertise sharing initiatives, such as this parasitology database, provide HCI-based collaboration opportunities for learners, educators, researchers, and practitioners.

In the departments of Small and Large Animal Clinical Sciences, a number of TEL projects have been developed including interactive, self-learning CD-ROMs, and DVDs, developed as alternatives to traditional laboratory manuals and live-animal teaching demonstrations. To date, this resource set includes Passing a Nasogastric Tube in the Horse, Canine and Feline Medical Exercises, Medical Imaging, and Basic Anesthetic and Surgical Principles. Equine Medicine and Canine Orthopedics resources are in development, and a proposal for the creation of a case-based interactive electronic learning environment for bovine disease diagnosis has been funded. In addition to meeting the needs of undergraduate students, the clinical science CAL resources have been designed to support learners enrolled in other veterinary colleges, the veterinary technology program at the Saskatchewan Institute of Applied Science and Technology, for continuing education credit options for practitioners and foreign graduates seeking practicing licenses in Canada, as well for workers in the beef-ranching community. Many of the completed instructional resources have been the focus of educational effectiveness research projects.

A major goal of many of the clinical science CAL projects is to reduce the number of live
animals used in teaching, thus promoting animal welfare, while continuing to provide highly relevant clinical material. To confirm that the approaches to creating and implementing these CAL modules were educationally sound, project members adopted a stance of research-based educational practice. Team members investigated current literature on veterinary education, active learning, and problem-based learning. While problem-based and active learning have become buzzwords in teaching-oriented journals (Hines et al., 2005), the instructional designers and WCVM faculty (Naylor, 2005) specifically employed Dale’s active learning model (see Figure 1).

Learning effectiveness studies suggest that knowledge retention improves with the number of senses stimulated and the degree of active involvement in the learning process (Dale, 1969; Peal & Wilson, 2001). Older methods of teaching often score poorly on the active-passive scale. Traditional lectures, where the professor verbally recites a synopsis of facts that the student copies, stimulate few senses and provoke little mental involvement. The degree of complexity in this task can be increased if the student has to mentally summarize the information before transcribing; however, new technologies make more effective learning methods possible. Computer-based presentations provide a method for accessing and displaying different types of material so that a variety of sensual stimuli can be used to improve retention. In addition, computer-based presentations offer a means of establishing relevance by incorporating actual recorded clinical case material into the presentation. Another benefit of computer-based models is the ability to add specific active learning exercises, which challenge the student, focus the learning, and give immediate feedback.

Accurate characterization of visual, auditory, or tactile phenomena is difficult using purely verbal or text-based descriptions. With paper-based books, this can be partially rectified by inserting annotated diagrams or images, but traditional books cannot reproduce sound or movement. As a result, meaning is lost, and learning may be misdirected. For example, an evaluation of the ability of veterinarians to describe and interpret common equine heart sounds found a lack of a common vocabulary, but not a lack of vocabulary.

**Figure 1. Active learning model (Adapted from Dale, 1969)**

![Active Learning Model Diagram](image-url)
In other words, wide varieties of words were used in an inconsistent and sometimes contradictory fashion (Naylor, Wolker, & Pharr, 2003); in addition, there was a lack of interpretive skills when the clinicians were played recordings of heart sounds from horses with common clinical arrhythmias (Naylor, Yadernuck, Pharr, & Ashburner, 2001).

Text, audio, still photos, medical illustrations, and video demonstrations are incorporated into the clinical science CAL projects. Active learning is stimulated with interactive case simulations or technical problem solving exercises. Veterinary students are asked to work through these materials and complete formative evaluation exercises before performing these techniques on live animals in the teaching hospital. This is regarded as both an example of good learning practice and of attention to animal welfare.

The emergent pedagogical foci on problem-based and active learning were also considered in some modules. This was done by first presenting a clinical case, accompanied by the usual “ill structured” problems, “similar to tasks a student will face in real world” veterinary practice (Schoenfeld-Tacher et al., 2005). Student learning was supported by access to textual and visual information describing the indications and contraindications of a specific technique, as well as demonstrations of common errors and their consequences.

Research-Based CAL Development: Equine

A self-learning computer module, How to pass a nasogastric tube in the horse, was developed by a team of veterinarians, along with instructional designers, and multimedia specialists. The computer-based learning module includes sections on indications for the technique, needed equipment, relevant anatomy, detailed instructions for performing the procedure, and common errors and complications. Each section consists of learning objectives and instructional text, supplemented with a combination of synchronized video clips of endoscopic and external camera views, audio, still photographs, and medical illustrations. The module concludes with an electronically scored student self-test, in which test questions are based upon both text and images (Naylor & Abutarbush, 2004).

The development of this module occurred over a two-year time-period. First, a WCVM faculty member and a resident (clinical graduate student) conducted an extensive literature review and clinical review of anatomy. From this basis, an initial project concept was developed. An exploration of ways to teach nasogastric intubation (NGI) resulted in the selection of synchronous video endoscopy of the internal and external passage of the stomach tube in living horses and dead horses, as well as photography of clinical specimens and medical illustrations.

A clinical resident, with the support of a WCVM faculty member, developed the first draft of the written content for the module. At this point, an instructional designer was added to the team to construct a project blueprint. In collaboration with media specialists, the team expanded the instructional design blueprint into a detailed storyboard, which identified text, visual, audio, and video components and sequences, as well as interactive elements. Illustrations were collaboratively developed by the faculty member, the resident, and a medical illustrator. Instructional objectives were aligned with student performance expectations and evaluation techniques. Media specialists collaborated with the team to select appropriate technologies, add to the overall concept of the project, gather, compile, and edit the video clips, and produce the visual design and navigation of the CAL learning environment. A multimedia programmer used Macromedia Director MX and Macromedia Flash MX to bring each of these elements together in an interactive
Integrating Human Computer Interaction in Veterinary Medicine Curricula

Throughout this process, the instructional designer tracked the overall progress of the project, adjusted the project blueprint and milestones, as required, and distributed updated documentation to group members.

Prior to the release of the CD-ROM for student use, a double blind, monocentric study involving 52 third-year student-participants in the Doctor of Veterinary Medicine program was conducted. The objective of this study was to evaluate the effectiveness of the NGI self-learning computer module compared with the effectiveness of traditional demonstration-based instruction (Abutarbush, Parchoma, Petrie, & Naylor, 2004). Participants were randomly assigned to control (traditional demonstration) and experimental (CAL) groups. Both groups received the same amount of instructional time.

In the NGI study, quantitative and qualitative data were collected and analyzed. Quantitative data were collected via student self-reports on a 5-point Likert scale (5-strong preference for CAL to 1-strong preference for traditional instruction). Twenty-six percent of participants reported either a preference or a strong preference for traditional instruction; 26% reported a preference for CAL; and 48% reported no preference. A comparison of the groups showed that within the control group the mean preference for CAL was slightly lower (2.57) than in the CAL group (3.29). However, the independent samples t-test showed no significant difference in preference between groups.

Learner confidence in mode of instruction received was measured using three questions. The overall quality of instruction was rated as high by 42.6% of all participants and very high by 57.4%. There was a mean rating of 4.6 on the 5-point scale for organization of instruction in the control group and 4.5 in the experimental group. High (46.8%) and very high (57.4%) levels of student-participant confidence in their preparedness to perform the naseogastric intubation procedure were reported. The group means for their level of confidence were 4.39 and 4.67 on the 5-point scale for the control and CAL groups respectively. Again, t-test results showed no significant difference between groups.

Encouragingly, the audio and visual clips on the CD were rated as being significantly better than the live horse demonstration in helping students learn to perform the procedure. Median values were 4 and 5 for the control and CAL groups respectively.

Participant achievement levels were measured by student scores on a knowledge quiz and during practical evaluations of performance of the technique. Each student-participant performance of the technique was videotaped from two camera angles. Two evaluators reviewed the videotapes and assessed student performances. Proficiency of technique performances was measured using an evaluation rubric (see Appendix A). Evaluators had no knowledge of which group individuals had been assigned.

Data collected from the knowledge quiz results and evaluation rubric were analyzed using descriptive statistics, non-parametric tests: the Wilcoxon Rank Sum and Chi-square, with the help of a computerized statistical package, Student Statistix 7 (Abutarbush et al., 2004). Quantitative results showed students in the CAL group performed better on the test of knowledge (median scores out of 10 were 9.67 for the CAL group and 8.1 for traditional demonstration group; analysis of significant difference resulted in P<0.001) (Abutarbush et al., 2004). In the assessment of the students’ practical ability, there was no significant difference in the number of attempts or number of times assistance was required during the performance of the procedure. The number of technical errors between the two groups was found to be consistent by both evaluators. However, the students in the CAL group needed significantly less time to perform the procedure than did the traditionally instructed group (Abutarbush et al., 2004). A reduction in time to complete the
performance of the procedure reduces stress to the horses, and therefore, contributes to improved animal welfare.

Following the experiment, student comments on their experiences in the evaluative study were gathered via conducting two semi-structured focus group debriefing sessions: one for volunteers from the control group, and one for volunteers from the experimental group. The focus groups were facilitated by the instructional designer, rather the faculty from the WCVM in order to ensure students felt able to comment freely, without concern for faculty members’ responses to individual comments.

Focus group transcripts were made anonymous, coded, and analyzed prior to results being shared with faculty members. The results of these focus group sessions contributed to refinements of the NGI CAL module, as well as to subsequent educational research directions.

As a follow-up to the pilot use of the NGI resource, a faculty member who was not yet involved in either the CAL development or the associated research study was interviewed after using the CD-ROM to teach the technique. A summary of the interview follows:

As a new faculty member at WCVM, I had not participated in development of the CAL module and had not taught the laboratory previously. I had reviewed the module prior to teaching the laboratory and compared it to my own experiences with the procedure.

I purposely limited instruction to the students reviewing the CD module without additional verbal instruction or demonstration of the procedure. It was intended to have each student review the CD individually; however, due to technical problems in the computer laboratory, several students had to share computers and review the CD as a group. Following review of the CD, I only provided logistical information, such as how many students were to work on any one horse, and gave instructions pertaining to safety measures. The students then proceeded to pass nasogastric tubes, with instructors present to answer questions and provide help as needed.

The students reviewed the CD module very willingly and completed most of the module in the time given. Most students had not reviewed the CD module prior to the lab. Several students volunteered positive feedback pertaining to the quality of the module. It was interesting to observe students reviewing the module as a group or individually; while students reviewing the module individually tended to complete it step by step, much discussion arose in the groups sharing one module and students explained to each other what was unclear to them and highlighted important points.

During the practical portion of the laboratory, all students were able to pass a nasogastric tube successfully in the time allotted, and no major problems, such as significant bleeding in the teaching horses arose. I was pleasantly surprised by this outcome as I had expected problems and remembered my frustration of not being able to complete the procedure as a veterinary student myself. The students approached the procedure without much hesitation and appeared to recall most of the instructions given on the CD. Help was needed mostly when problems arose with restraint of horses, or to reassure students that they had completed the procedure successfully.

It was again very interesting to observe interactions between students. Five to six students worked on one horse and took turns completing the procedure. Students who had already completed the procedure could overhear recounting their experiences and providing advice to their peers. The quality of their explanations was as good, if not better, than that I could have given myself, especially as students could relate to the initial insecurities much better than a more experienced person. Following the laboratory, several students asked to check out the CD module for additional review.
This was my first experience using a CAL module for student instruction and it was a very positive experience. I have since become involved with development of additional modules to teach equine procedures, and am looking forward to using them in future medical exercises laboratories.

Documentation of the learning benefits of CAL in the NGI study and continuing student and faculty enthusiasm for the use of the NGI resource prompted faculty interest in developing more CAL modules. The Department of Large Animal Clinical Science has undertaken a new and more ambitious project, Techniques in Equine Medicine. This project is slated to include modules on performing a number of diagnostic techniques including transtracheal wash, pleurocentesis, liver biopsy, abdominocentesis, urine collection, and cerebrospinal fluid collection from atlanto-occipital and lumbosacral sites.

The Techniques in Equine Medicine project completion date is projected for 2007. While this project is much greater in scope than the NGI project, the lessons learned in the NGI development, as well as elements, such as the learning environment’s structures, functionalities, and visual design will be reused, thus creating time and cost efficiencies.

Research-Based CAL Development: Canine and Feline

A self-learning computer module, Canine and Feline Medical Exercises, was developed by a veterinarian at the WCVM who was not involved in the NGI module in conjunction with a University of Saskatchewan team of instructional designers and media specialists. The development of this project took place over the course of three years, and an educational study is underway to evaluate the modules. The canine-feline medical exercises CAL modules were structured similarly to the NGI design, but a series of adaptations were made to accommodate multiple exercises. Five systems were identified (respiratory, gastrointestinal, urinary, dermatology, joints), and the important diagnostic and sample collection techniques taught to veterinary students for each system were determined.

Each of the 28 technique modules was designed to include the following components:

- A list of indications and contraindications for performing the technique
- A link to a case description to illustrate where the technique was successful in obtaining a diagnosis or influencing therapy
- A list and photographs of the specialized equipment necessary to perform the technique
- A step-by-step written description of the technique paired with digital photographs (with anatomic landmarks labeled) and/or drawings illustrating each important step of the procedure; the written description was then followed by a narrated videotape of the procedure as it was performed, re-emphasizing the important points already discussed
- A description of appropriate sample handling, submission for analysis, and (where appropriate) interpretation of results
- A list of possible complications of the technique
- Self-tests emphasizing key anatomic landmarks or details of the techniques

A detailed list of the digital still photos and video requirements needed was prepared for each technique. Instructional designers formatted all of this information into a course design map. This map clearly defined the learning objectives, and the content, activities, and media requirements for each objective. The design map was then circulated among the development team for approval.

Following the development of the design map, the faculty member met several times with an in-
structional designer to blueprint each procedure. The blueprint provided the media team with the information it needed to produce the final product. Each screen of the final product was sketched. The media team knew exactly what was required to appear on each screen and the written text to accompany it. The completed blueprints were sent to the media team and the design team.

Video footage and still digital photos were taken during two days of demonstrations of the techniques. Once the raw material had been gathered, the media team began to edit and assemble the pieces. Errors and omissions were discovered and corrected. A multimedia programmer used Macromedia Director MX and Macromedia Flash MX to bring each of these elements together in an interactive learning experience.

Five fourth-year veterinary students and two veterinary technicians, under the guidance of an instructional designer, user-tested the set of two CD-ROMs. User testing resulted in a series of revisions and improvements to the modules. These revisions included additional feedback to learners on their performance in interactive exercises and self-tests.

As an extension of the NGI research study into the educational effectiveness of WCVM’s CAL resources, the research study into the Canine and Feline Medical Exercises modules will investigate student learning styles, student preferences, and module effectiveness as a learning tool. This direction has been taken, in part, because student-participants in the NGI study reported individual learning style as a variable in learner preferences for traditional or CAL-based instruction. A literature review was undertaken, and peer-reviewed articles that support the student reports were identified (Buchanan, Carter, Cowgill, Hurley, Lewis, MacLeod, Melton, Moore, Pessah, Roberson, Robertson, Smith, & Vadenplas, 2005; Dale, Nasir, & Sullivan, 2005; Schoenfeld-Tacher, McConnell, & Schultheiss, 2003). The Canine-Feline study compares traditional versus CAL instruction. In addition, research team members want additional insight into the effect of learning styles on student responses to CAL resources and will attempt to discover if individual learning-style preferences are a significant variable in instructional mode preference or in student achievement.

Fifty-eight third-year veterinary students agreed to participate in the Canine-Feline research study. The research design included quantitative and qualitative research methodologies. The research team assigned all participants numbered codes to insert in the identity portion of the online version of the Felder-Silverman Index of Learning Styles (ILS) measurement instrument, and all student participants completed the ILS questionnaire to establish their learning style. Participants were assigned to experimental (CAL) and control (traditional laboratory manual) groups using block randomization. Participants were divided into four groups based on their learning styles and then randomized within groups. No member of the faculty of Veterinary Medicine knew which students or codes had been assigned to control or experimental groups.

The ILS measurement instrument has been rigorously tested via multiple studies across post-secondary disciplines and institutions for reliability and validity. Reliability and validity results have been published in peer-reviewed, scholarly publications (Litzinger, Lee, Wise, & Felder, 2005). Results of the ILS measurement instrument are only considered valid and reliable when the instrument is used to help instructors “to achieve balanced [learning-style] course instruction” and to “help learners understand their learning strengths and areas for improvement” (Felder & Spurlin, 2005, pp. 110-111). In this study, the researchers will use participant scores on the ILS for the purpose of seeking further information on the comparative effectiveness of traditional and computer-assisted modes of instruction for students with different learning styles as well as evaluating the impact of learning styles on
learning-mode preference. It is anticipated that the results of this study will have an impact on the design and development of future products to support student learning in HCI environments.

Participants were asked to complete a short survey, including mode-of-instruction-preference questions with responses selections on a Likert scale. Likert scale results will be compared to ILS measurement results in order to investigate potential linkages between learning styles and learning mode preferences. The correlation between learning styles and learning mode of instruction preference will be examined using Spearman’s correlation coefficient.

The association between learning styles and achievement will be assessed using regression analysis. Learning styles will be measured along the visual versus verbal continua (Felder & Spurlin, 2005) in order to determine the degree of polarity (or lack of polarity) of individual visual versus verbal learning styles within the participant group. Data analysis will include measures of potential combined effects of self-reported learning preferences and ILS scores. Achievement will be measured using student performance on a standardized practical oral examination.

The question of whether multimodal instruction using the CAL modules results in better achievement in students of particular learning styles will be assessed via regression analysis. The study will specifically examine whether use of the CAL modules or the traditional laboratory manual worked differently for students with different learning styles. Data will be analyzed assessing whether or to what extent instruction modules designed to support individual learning styles might interact with learner achievement.

The results of the Canine and Feline Medical Exercises CAL study will post-date the publication of this chapter. The research team hopes to use the evidence gathered to identify ways to improve future instructional design and development to better meet WCVM learner needs.

CONCLUSION

Given the significant challenges and pressures influencing veterinary colleges, teaching hospitals, and educators, new strategies for providing high quality learning must be considered. CAL is an effective tool that can be used to enhance the veterinary curriculum and provide increased student access to clinical material while reducing the unnecessary use of live animals in teaching. The CAL-based veterinary medical learning resources developed at the WCVM represent a collaborative effort between the veterinary clinicians and researchers who conceived each project and served as content experts and the instructional designers, media specialists, and multimedia programmers who fine-tuned each project as it was developed. The results of educational research performed on the CAL modules as they have been implemented in the curriculum have already directly influenced the development of future projects integrating HCI into the WCVM curriculum and future directions for research focused on these projects. It is hoped these efforts will contribute to the globalization of the veterinary profession through evidence-based approaches to innovation in the use of HCI in veterinary education.

ACKNOWLEDGMENTS

Research and development projects described in this chapter have been supported by the Government of Saskatchewan’s Technology Enhanced Learning (TEL) initiative, and the Association for Media and Technology in Education in Canada (AMTEC) Trust.

Lily McCaig, a WCVM student, contributed to the literature review and the development of the equine medicine CAL modules.

Juliana Deubner, medical illustrator for the WCVM, prepared all medical images for the CAL learning modules, as well as the figure for the chapter.
Frank Bulk, Gabe Ng, and Wayne Giesbrecht, multimedia developers at the University of Saskatchewan, have contributed their talents to many projects described in the chapter.

REFERENCES


Integrating Human Computer Interaction in Veterinary Medicine Curricula


Integrating Human Computer Interaction in Veterinary Medicine Curricula


APPENDIX A: RUBRIC USED FOR EVALUATION OF STUDENT PERFORMANCE OF NASOGASTRIC INTUBATION

Student # ______________
Tape # / Angle ______/______ (overhead / floor views of camera angle and tape number)
Time frame on tape # _____ / _____ /______ to _____ /_____/______ (OH or FL time-stamp: circle your initials)

<table>
<thead>
<tr>
<th>Time taken to complete the procedure</th>
<th>minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of times the student asked for assistance (circle one below)</td>
<td>0</td>
</tr>
</tbody>
</table>

1. Technical Errors in NGI procedure
   - Failed to evert the false nostril
   - Occluded the nostril
   - Left the free end of the tube unattended
   - Incorrect restraint
   - Failed to do measurements
   - Failed to lubricate the tube
   - Failed to stimulate swallowing
   - Failed to kink the tube
   - Bleeding occurred with tube passage
   - Failed to flex the neck/left the neck stretched
   - Failure to perform 3 checks of tube position

2. Tube position checking procedures
   - Ensured correct resistance on the tube
   - Checked whether air could be sucked back
   - Checked whether the tube could be seen moving on the outside of the neck
   - Inflated the esophagus
   - Rattled trachea to see if can feel the tube
   - Checked for sound or smell of stomach gas
   - Had an assistant use a stethoscope to listen for stomach sounds
   - Checked whether the tube could be felt moving on the outside of the neck
   - Coughing not elicited

3. Number of times to successful completion

---

This work was previously published in Enhancing Learning Through Human Computer Interaction, edited by E. McKay, pp. 204-220, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
INTRODUCTION

A portal is defined as an entrance point to online content. The portal concept has evolved across a number of markets and applications. Customer portals focus on individual customer and offer a one-stop Internet access. By providing a number of services, such as searches, shopping, e-mail, and games, portals allow individuals to avoid browsing the Web but to in-fact rely and stay at one Web site like a one-stop shop. Accordingly, portals drive eyeballs, and hence create and drive advertising revenue and alliances. The concept of a single public port to given content on the Internet is used as a means of pulling in a large number of users. As an example, America Online (AOL) acts as a portal site to general Web content. It is a specialized portal created by AOL and also has content from partners such as Time Warner (Kleindl, 2003). This article reviews the role of portals in consumer search behavior and certain aspects in marketing.

PORTALS AND PRODUCT CUSTOMISATION

A key function of marketing is to match buyers and sellers, and facilitate transactions; to do this a firm needs to create the proper institutional infrastructure. It has been found that digital information goods, such as news articles, digital images, or music allow perfect copies to be created and distributed almost without cost via the Internet. With the introduction of the Internet as a commercial medium for businesses to conduct their activities, various studies have found that the technology is leading to aggregation. This, in turn, is fast becoming a profitable strategy for marketers, as the marginal production costs are low and consumers are generally homogenous. Several Internet-based technologies assist buyers searching: multimedia, high bandwidth, and rating sites provide more product information. These search engines can be hierarchical directories like Yahoo, generic tools like Alta Vista (in
early 1998) or specialized tools that work best in the context of specific markets like Pricewatch, ComputerESP for computers, or Expedia and Travelocity for travel (Casagranda, Nicholas, & Stevens, 1998).

Customer portals should provide companyspecific information for customers, such as product information, inventory and order tracking, help desk applications, and other services (Kleindl, 2003). Marketers should begin considering portals as the brains of the organization as they can provide employees with vital information for success in hyper-competitive marketplace, in turn can secure the survival of the organization. The method is cost-effective because portal technology uses artificial agents, tiny programs to find and organize information rather than salaried employees.

Clarke, III and Flaherty (2003) suggest that portals are the most valuable land on the Web. According to them, about 90% of Internet traffic goes to 10% of Web sites, among which portals are the largest shareholders of that traffic. The authors have also found that about 15% of all Web page-view traffic goes through the top nine portals. Hence, this heavy traffic flow creates a unique position for portals as part of the overall marketing strategy of all organizations.

Some suggest that with portal technology it is possible for an individual to buy a newspaper at a local newsagent and this newspaper can be tailored to suit the person’s specific information needs. This newspaper can contain a section on industry news, another on company news, and a third on all financial reports, all of this information may be very relevant to the person. If such a newspaper could be economically produced the reader would not need to buy a whole newspaper to read but just a few pages. Such customization can be achieved economically with portal technology because the artificial agents used in portals are programmed to search and index sites containing information the user specifies as relevant (Kotorov, 2001).

Slywotzky (2000) extends this concept of customization of products and services using portal technology to newer heights. According to the author, customers will soon be able to describe exactly what they want, and suppliers will be able to deliver the desired product or service without compromises or delays. This innovation is what the author calls “choiceboard,” this concept includes interactive online systems that allow individual customers to design their own products by choosing from a menu of attributes, components, prices, and delivery options. The role of the customer in this system shifts from passive recipient to active designer. The shift is just the most recent stage in the long-term evolution of the customers’ roles in the economy.

It was further illustrated that with a choiceboard system, marketers will see a major shift of customers becoming product makers rather than product takers. Traditionally, companies create fixed product lines that represent their best guesses about what buyers will want, and buyers make do with what they are offered. There may be some minor tailoring at the point of purchase—a few optional features or add-ons—but by and large the set of choices is fixed by long before customers even begin to shop (Slywotzky, 2006).

The choiceboard concept became an interactive, online system model, allowing individual customers to design their own products by choosing from a menu of attributes, components, prices, and delivery options. The customers’ selections send signals to the supplier’s manufacturing system that set in motion the wheels of procurement, assembly, and delivery. They are already in use for example; customers can design their own computers with Dell’s online configurator. They can create their own dolls with Mattel’s My Design Barbie, assemble their own investment portfolios with Schwab’s mutual fund evaluator, and even design their own golf clubs with Chipshot.com’s PerfectFit system. This Choiceboard is still in its infancy, as it is involved in less than 1% of the $30 billion world economy (Slywotzky, 2006).
By providing a number of services, such as searches, shopping, e-mail, and games, portals allow individuals to avoid browsing different other Web sites, but to stay at one single portal type site. Since the site drives eyeballs, it in turn will drive advertising revenue and alliances. The concept of a single public port to access content is used as a means of pulling in a large number of users (Kliendl, 2003).

CONSUMER BEHAVIOR AT PORTALS

The growth of the Internet and its immense capability of providing consumers with product and service information has empowered the consumer immensely. Consumers are becoming more mature, sophisticated, and intelligent. These days they are seeking a higher levels of product information before making purchasing decisions. The rapid advancements in Web technology have enhanced consumer’s decision-making outcomes. The creation and subsequent growth of software and technological devices such as smart agents that are linked to portals have provided an intelligent interface for the consumer. These computer decision aids improve transactional efficiency by providing merchandising and sales information to consumers, offering sales support, and facilitating sales promotions, while at the same time, enhancing the consistency, availability, and quality of support to consumers.

In a study to test the relationship between the use of these smart agents, or query-based decision aids (QDBA) as they are referred to, and consumers, it was found that the greater the amount of relevant information the decision maker has, the greater is his or her confidence in judgement. The research study developed and tested a general model for understanding the influence of query-based decision aids on consumer decision making in the e-commerce environment. The results showed that the use of a well designed QDBA led to increased satisfaction with the decision process, and increased confidence in judgements. The research subjects who had access to QDBA perceived an increased cost saving and a lower cognitive decision effort associated with the purchase decision. The conclusion proved that subjects who had access to the QDBA, liked the interface, and had more confidence in their judgements in comparison to subjects who did not have access to QDBA (Pereira, 1999).

In their study, Meisel and Sullivan (2000), found that most Web surfers and shoppers want portals to conduct five important functions as follows:

• provide easy, convenient, and organized way for users to use the Internet;
• act as a filter and hence helping in the decision making process of the purchase online;
• assure users of the integrity of the sites for Web transactions;
• provide users access to propriety content and/or communication technologies like Internet telephony and e-mail; and
• finally, to facilitate the electronic equivalent of one-stop shopping for the user.

Studies have indicated that the main reason individual’s use portals is for gathering information, these fall into two categories namely: personal needs, covering leisure (sport, films, games, specific niche hobbies, chat) medical information, news and politics, local community and historical information; and information gathering, which include the gathering of information for business needs, this can cover technical resource information, academic research and company information. Portals support the information search stage of the buying process; research has found that consumers do make use of portals for the decision-making process in consumption behavior (Michael, 2006).

Hanson (2000) found that most Web users start their online activities at one of the main search
or directory portal sites, hence making portals an important source of traffic that can be obtained for free. Managing an organization’s portal presence requires traffic-building efforts that combine strategic and tactical activities. A key strategic initiative to manage ones portal presence is to classifying a site carefully using proper keywords, descriptors, and categories. This is very important especially for directories that group sites into specific classification systems.

Marketers of portals should work with the directory personnel to make sure that the latter correctly locate the company’s site to provide a steady stream of visitors. Hanson (2000) further suggests that there needs to be a continuous tactical attention to effectively leverage the portals, especially search engines. He states that consumers search using a range of methods these could include things like keywords in search engines, meta-tags, and various other links. These variables should then be kept in mind by marketers and be used strategically with search engines to enable it (search engine) to retrieve proper results for the searcher/surfer. A Web site manager must monitor and improve the chances of material being found and retrieved early in the list of results of these pages.

**CONCLUSION**

Portals as the definition suggests are gateways to the Internet, they should be used as strategic tools in the marketing process. Marketers need to keep abreast as to the growth, potential, and changing nature of these sites which play a key introductory role to Web searchers. They are best summed up as very large aggregators that will become more and more of a one-stop shop for consumers. Portals were one of the first pure e-commerce type companies to focus and create online brands, true examples of these are the popularity of brands such as Yahoo!, Alta Vista, Amazon, Travelocity, and the likes.

It has been found that consumers rely on branded names especially in this mire of products and services that is available over the Internet. Research has also proved that if in doubt, consumers are straightaway attracted to the online brands that are become familiar with, little wonder that Amazon is supposedly the most successful pure online company and brand. Portals have matured to become a key trading exchange intermediary between consumers and businesses, and also between business and other businesses. They (portals) recent focus is now on convenience, price, and variety. In their role as business to business exchanges, portals are rapidly taking the form of creating strategic alliances between like minded companies. It now seems rest assured that portals the gateways will be the key in our future cyber journey.

**REFERENCES**


**KEY TERMS**

**Choiceboard**: This concept includes interactive online systems that allow individual customers to design their own products.

**Directory**: The word directory is used in computing and telephony meaning a repository or database of information.

**Marketing**: A function within an organization, of a set of processes for creating, communicating and delivering value to customers, to benefit the organization and its stakeholders.

**Online**: A term used to describe information that is accessible through the Internet.

**Portal**: An entrance point to online content.

**Search Engine**: A search engine or search service is a program designed to help find information stored on a computer system such as the World Wide Web, inside a corporate or proprietary network or a personal computer.

**Web Surfers**: Consumers who go online, the phrase “surfing the Internet” was first popularized in print by librarian Jean Armour Polly in an article called *Surfing the INTERNET*, published in the Wilson Library Bulletin in June, 1992.

*This work was previously published in Encyclopedia of Portal Technologies and Applications, edited by A. Tatnall, pp. 811-813, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 4.18
Human Factors in the Development of Trend Detection and Tracking Techniques

Chaomei Chen
Drexel University, USA

Kaushal Toprani
Drexel University, USA

Natasha Lobo
Drexel University, USA

INTRODUCTION

Trend detection has been studied by researchers in many fields, such as statistics, economy, finance, information science, and computer science (Basseville & Nikiforov, 1993; Chen, 2004; Del Negro, 2001). Trend detection studies can be divided into two broad categories. At technical levels, the focus is on detecting and tracking emerging trends based on dedicated algorithms; at decision making and management levels, the focus is on the process in which algorithmically identified temporal patterns can be translated into elements of a decision making process. Much of the work is concentrated in the first category, primarily focusing on the efficiency and effectiveness from an algorithmic perspective. In contrast, relatively fewer studies in the literature have addressed the role of human perceptual and cognitive system in interpreting and utilizing algorithmically detected trends and changes in their own working environments. In particular, human factors have not been adequately taken into account; trend detection and tracking, especially in text document processing and more recent emerging application areas, has not been studied as integral part of decision-making and related activities. However, rapidly growing tech-
nology, and research in the field of human-computer interaction has opened vast and, certainly, thought-provoking possibilities for incorporating usability and heuristic design into the areas of trend detection and tracking.

**BACKGROUND**

In this section, we briefly review trend detection and its dependence on time and context, topic detection and tracking, supported by instances of their impact in diverse fields, and the emerging trend detection especially for text data.

**Trend Detection**

A trend is typically defined as a continuous change of a variable over a period of time, for example, unemployment numbers increase as the economy enters a cycle of recession. Trend detection, in general, and topic detection techniques are groups of algorithmic tools designated to identify significant changes of quantitative metrics of underlying phenomena. The goal of detection is to enable users to identify the presence of such trends based on a spectrum of monitored variables. The response time of a detection technique can be measured by the time duration of the available input data and the identifiable trend; it is dependent on specific application domains. For example, anti-terrorism and national security may require highly responsive trend detection and change detection capabilities, whereas geological and astronomical applications require long-range detection tools. Other applications of this technology exist in the fields of business and medicine.

Much research has been done in the field of information retrieval, automatically grouping (clustering) documents, performing automated text summarization, and automatically labeling groups of documents.

Policy makers and investigators are, obviously, eager to know if there are ways that can reliably predict each turn in the economy. Economists have developed a wide variety of techniques to detect and monitor changes in economic activities. The concept of business cycles is defined as fluctuations in the aggregate economic activities of a nation. A business cycle includes a period of expansion, followed by recessions, contractions, and revivals. Three important characteristics are used when identifying a recession: duration, depth, and diffusion — the three Ds. A recession has to be long enough, from a year to 10 years; a recession has to be bad enough, involving a substantial decline in output; and a recession has to be broad enough, affecting several sectors of the economy.

**Topic Detection and Tracking**

*Topic Detection and Tracking* (TDT) is a sub-field primarily rooted in information retrieval. TDT aims to develop and evaluate technologies required to segment, detect, and track topical information in a stream consisting of news stories. TDT has five major task groups: (1) story segmentation, (2) topic detection, (3) topic tracking, (4) first story detection, and (5) story link detection. Topic detection focuses on discovering previously unseen topics, whereas topic tracking focuses on monitoring stories known to a TDT system. First story detection (FSD) aims to detect the first appearance of a new story in a time series of news associated with an event. Roy, Gevry, and Pottinger (2002) presented methodologies for trend detection. Kontostathis, Galitsky, Roy, Pottinger, and Phelps (2003) gave a comprehensive survey of emerging trend detection in textual data mining in terms of four distinct aspects: (1) input data and attributes, (2) learning algorithms, (3) visualization, and (4) evaluation.
TDT projects typically test their systems on TDT data sets, which contain news stories and event descriptors. The assessment of the performance of a TDT algorithm is based on Relevance Judgment, which indicates the relevancy between a story and an event. Take the event descriptor Oklahoma City Bombing as an example. If a matching story is about survivors’ reaction after the bombing, the relevance judgment would be Yes. In contrast, the relevance judgment of the same story and a different event descriptor U.S. Terrorism Response would be No. Swan and Allan (1999) reported their work on extracting significant time varying features from text based on this type of data.

An interesting observation of news stories is that events are often reported in burst. Yang, Pierce, and Carbonell (1998) depicted a daily histogram of story counts over time. News stories about the same event tend to appear within a very narrow time frame. The gap between two bursts can be used to discriminate distinct events.

Kleinberg (2002) developed a burst detection algorithm and applied to the arrivals of e-mail and words used in titles of articles. Kleinberg was motivated by the need to filter his e-mail. He expected that whenever an important event occurs or is about to occur, there should be a sharp increase of certain words that characterize the event. He called such sharp increases bursts. Essentially, Kleinberg’s burst detection algorithm analyzes the rate of increase of word frequencies and identifies the most rapidly growing words. He tested his algorithm on the full text of all the State of the Union addresses since 1790. The burst detection algorithm identified important events occurring at the time of some of the speeches. For example, depression and recovery were bursty words in 1930-1937, fighting and Japanese were bursty in 1942-1945, and atomic was the buzz word in 1947 and 1959.

**EMERGING TREND DETECTION (ETD)**

**ETD for Text Data**

Unlike financial and statistical data typically found in an economist’s trend detection portfolio, ETD in computer science often refers to the detection of trends in textual data, such as a collection of text documents and a stream of news feed. ETD takes a large collection of textual data as input and identifies topic areas that are previously unseen or are growing in importance with in the corpus (Kontostathis et al., 2003). This type of data mining can be instrumental in supporting the discovery of emerging trends within an industry and improving the understanding of large volumes of information maintained by organizations (Aldana, 2000). In the past few years, many companies have been storing their data electronically. As the volume of data grows, it will hold information, which if analyzed in the form of trends and patterns, can be valuable to the company, provided it is appropriately and accurately extracted. By using ETD, companies can extract the meaningful data and use it to gain a competitive advantage (Aldana, 2000). ETD provides a viable way to analyze the evolution of a field. The problem switches from analyzing huge amounts of data to how to analyze huge amounts of data.

**The Role of HCl in ETD**

ETD systems are complicated to make and understand, thus there are many HCI issues that must be considered. First of all, the system should let the user define what an emerging trend is. In general, an emerging trend can be defined as a significant quantitative growth over time. However, what counts as significant should not be entirely determined by computer algorithms.
Many ETD algorithms have different threshold levels to define a topic as an emerging trend. Thus threshold levels should not be fixed and unchangeable for a system. Also, the user should be able to define what documents are in the data corpus. Additionally, the algorithm should be hidden from the user. Ideally, the system would take its inputs and produce the outputs. When the user is given information, pertaining to inputs and outputs, sufficient amounts of user guidance should be provided. The design of an ideal user interface of a computer-based information system should be intuitive and self-explanatory. Users should feel that they are in control and they can understand what is going on. Despite the technical complexity of an underlying algorithm, the user interface should clearly convey the functions to the user (Norman, 1998).

Once a new trend is found, the system should include some mechanisms to define the essence of the new trend. A text summarization algorithm is a possible solution to this problem. Text summarization is capturing the essence of a data set (a single paragraph, document, or cluster) after reviewing the entire data set and producing output that describes the data set.

Once the data corpus is scanned, the user should be provided with feedback about the corpus. The user should be provided with information like the number of documents found, number of topics (or trends) found, number of new trends found, and other related information. For example, the system studied by Masao and Kôiti (2000), produces an entity-relationship (ER) graph showing the relation of topics. This not only shows the user what new trends were found, but also shows how they are related. ETD systems should also support an adaptive search mechanism. Users should have the option of providing keywords to search for emerging trends in specific fields.

Applications

Automatic trend detection has benefited a wide range of applications. An analyst will find emerging trend detection techniques useful in his area of work. The most generic application is to detect a new topic in a field and track its growth and use over time (Roy et al., 2002). Two examples are cited in the following sections.

European Monitoring Center for Drugs and Drug Addiction (EMCDDA)

The EMCDDA was disappointed when it realized that it failed to recognize the emerging trend in the use of the drug ecstasy. “...earlier identification of new drug consumption patterns would allow more time to assess the likely impact of such changes and, therefore, facilitate the earlier development of appropriate responses” (EMCDDA, 1999). With an effective trend detection system, agencies like the EMCDDA can prepare for and prevent the associated problems with a drug epidemic. However, with the number of documents in some databases reaching over 100,000 a manual review of the data is impossible.

XML

The emergence of XML in the 1990s is shown in Figure 1 in terms of the growing number of articles published each year on the second-generation language of the World Wide Web. Market and field analysts will find such knowledge of an emerging trend particularly useful. For instance, a market-analyst watching a biotech firm will want to know about trends in the biotechnology field and how they affect companies in the field (Kontostathis et al., 2003).

Stock market analysts rely on patterns to observe market trends and make predictions. In
The general goal is to identify patterns from a corpus of data. In the past, analysts have relied on the human eye to discover these patterns. In the future, trend and topic tracking systems can take over this role, thus providing a more efficient and reliable method for stock market analysis.

**FUTURE TRENDS**

The future is promising for HCI concepts to be heavily embedded in the analysis and design of emerging trend detection (ETD) systems. Powerful data modeling techniques can make salient patterns clearer and in sharper contrast. Some of the major technical problems are how to make the changes over time easy to understand and how to preserve the overall context in which changes take place.

ThemeRiver is a visualization system that uses the metaphor of a river to depict thematic flows over time in a collection of documents (Havre, Hetzler, Whitney, & Nowell, 2002). The thematic changes are shown along a time line of corresponding external events. A thematic river consists of frequency streams of terms; the changing width of a stream over time indicates the changes of term occurrences. The occurrence of an external event may be followed by sudden changes of thematic strengths. On the one hand, searching for an abruptly widened thematic stream is a much more intuitive task to detect a new story than text-based TDT systems that can only report changes in terms of statistics.

There are many things to keep in mind while developing an HCI-friendly ETD system. The basic usability goals can be used as a guideline to producing a user-friendly ETD system. By striving to make a system learnable, efficient, memorable, keep errors under control, and give the user satisfaction from using the system, the foundation for an HCI friendly system is laid. Table 1 defines each of the usability goals and how they can be applied in an ETD system.
**Human Factors in the Development of Trend Detection and Tracking Techniques**

Table 1. Usability goals and how to apply them to ETD

<table>
<thead>
<tr>
<th>Usability Goal</th>
<th>Definition</th>
<th>ETD Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learnability</td>
<td>“How easy the system is to learn” (Rozanski &amp; Haake 2003)</td>
<td>The system must be easy to learn for people from a wide variety of fields, including those with non-technical backgrounds.</td>
</tr>
<tr>
<td>Efficiency</td>
<td>“How quickly users can complete their tasks” (Rozanski &amp; Haake 2003)</td>
<td>The system should let the user focus on issues that are relevant to trend detection, without having to worry about issues with the system.</td>
</tr>
<tr>
<td>Memorability</td>
<td>“How easy the system is to remember” (ROZANSKI &amp; HAAKE 2003)</td>
<td>Users should not have to relearn the system each time they want to use it.</td>
</tr>
<tr>
<td>Control of Errors</td>
<td>Prevention and recovery from errors (ROZANSKI &amp; HAAKE 2003)</td>
<td>The system design should make errors less likely to happen, and when they do happen, the system should help the user out of the errors.</td>
</tr>
<tr>
<td>User Satisfaction</td>
<td>“How much users like the system” (ROZANSKI &amp; HAAKE 2003)</td>
<td>The users should be able to accomplish their goals without frustration.</td>
</tr>
</tbody>
</table>

A set of usability heuristics, proposed by Jakob Nielson (n.d.), can also pose as a good rule of thumb (see Table 2).

Task analysis is a detailed description of an operator’s task, in terms of its components, to specify the detailed human activities involved, and their functional and temporal relationships (HCI Glossary, 2004). By having users describe their process, step-by-step designers can learn much about the user’s behavior. When conducting task analysis, have the users describe “the steps they would follow, the databases and tools they would use, and the decision points in the process” (Bartlett & Toms, 2003).

**CONCLUSION**

Emerging trend detection is a promising field that holds many applications. However, for ETD systems to reach their full potential, they must be effective, easy to learn, easy to understand, and easy to use. A poorly-designed system will shun users away from this technology. It is important to remember that ETD systems are interactive systems. An ETD system, that just takes a data corpus and scans it, is not an effective one. Users must be able to define and experiment with thresholds, view feedback about the data corpus, and be able to understand new trends.

**REFERENCES**


Bartlett, J. C., & Toms, E. G. (2003, July 28-August 1). Discovering and structuring information flow among bioinformatics resources. In the *Proceedings of the 26th annual international ACM SIGIR conference on Research and development*
Table 2. Nielsen’s heuristics and application to ETD

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visibility of system status</td>
<td>While algorithms may take a while to process, there should be feedback so the user knows the progress of the system. (Nielsen)</td>
</tr>
<tr>
<td>Match between real world and system</td>
<td>The system directions should be presented in language the user can understand; avoid complicated jargon and technical terms. (Nielsen)</td>
</tr>
<tr>
<td>User control and freedom</td>
<td>The user should be able to set the various thresholds that go into defining topics and emerging trends. (Nielsen)</td>
</tr>
<tr>
<td>Consistency and standards</td>
<td>Uniform color schemes and presentation of data are necessary. (Nielsen)</td>
</tr>
<tr>
<td>Error prevention</td>
<td>Steps should be taken to prevent users from entering thresholds that do not work and starting processes without sufficient input (Nielsen)</td>
</tr>
<tr>
<td>Recognition rather than recall</td>
<td>Users should not have to remember long, complicated processes. The directions should be presented to them on the screen, or the setup should give the users clues on what to do next (Nielsen)</td>
</tr>
<tr>
<td>Flexibility and efficiency of use</td>
<td>The user should be able to easily change thresholds to compare results. There should be shortcuts available for more experienced users as well (Nielsen)</td>
</tr>
<tr>
<td>Aesthetic and minimalist design</td>
<td>The interface should be kept simple (Nielsen)</td>
</tr>
<tr>
<td>Help users recognize, diagnose, and recover from</td>
<td>Errors should be presented in a manner so that it does not look like regular data (Nielsen)</td>
</tr>
<tr>
<td>errors</td>
<td></td>
</tr>
<tr>
<td>Help and documentation</td>
<td>Ample user manuals should be provided and should be presented in simple language (Nielsen)</td>
</tr>
</tbody>
</table>
**Human Factors in the Development of Trend Detection and Tracking Techniques**


**KEY TERMS**

**Burst Detection**: The identification of sharp changes in a time series of values. Examples of bursts include the increasing use of certain words in association with given events.

**Information Visualization**: A field of study aims to utilize human’s perceptual and cognitive abilities to enable and enhance our understanding of patterns and trends in complex and abstract information. Computer-generated 2- and 3-dimensional interactive graphical representations are among the most frequently used forms.
**Intellectual Turning Points:** Scientific work that has fundamentally changed the subsequence development in its field. Identifying intellectual turning points is one of the potentially beneficial areas of applications of trend detection techniques.

**Paradigm Shift:** A widely known model in philosophy of science proposed by Thomas Kuhn. Paradigm shift is regarded as the key mechanism that drives science. The core of science is the domination of a paradigm. Paradigm shift is necessary for a scientific revolution, which is how science advances.

**Topic Detection and Tracking:** A sub-field of information retrieval. The goal is to detect the first appearance of text that differs from a body of previously processed text, or to monitor the behaviour of some identified themes over time.

**Trend:** The continuous growth or decline of a variable over a period of time.

**Trend Detection:** Using quantitative methods to identify the presence of a trend. A number of domain-specific criteria may apply to determine what qualifies as a trend, for example, in terms of duration, diversity, and intensity. Primary quality measures of trend detection include sensitivity and accuracy.

**Turning Point:** A turning point marks the beginning or the end of a trend. For example, the point at which economy turns from recession to growth.
Chapter 4.19
The Internet, Health Information, and Managing Health: An Examination of Boomers and Seniors

Christopher G. Reddick
The University of Texas at San Antonio, USA

ABSTRACT
This article examines the use of the Internet for gathering health information by boomers and seniors. This study attempts to determine whether online health seekers (individuals that have Internet access and have searched for health information online) have changed their behaviors from the information they found online. Essentially, has online health information helped them to manage their health more effectively? This research analyzes the Kaiser Family Foundation e-Health and the Elderly public opinion dataset of access by boomers and seniors to online health information. The major results indicate that boomers marginally use online health information more than seniors for the management of their health. The most significant results indicated that boomers and seniors who are more aware and have positive feelings toward online health information would use it more to manage their health.

INTRODUCTION AND BACKGROUND
For baby boomers, the Internet has become the most important source of health information other than consultation with their family doctor (Kaiser Family Foundation, 2005). The focus of this article is on both baby boomers, those in the age range of 50 to 64, and seniors, or those 65 and older. This study examines the use of online health information by baby boomers and seniors and how they use the information for managing their health. The primary objectives of this article are to examine the differences in behavior between boomers and seniors and to test for the presence of a variety of associations among their characteristics and a number of management of health variables.

This study explores five specific questions. First, are there any differences between boomers and seniors and their access to health informa-
tion for managing health? Second, will healthier boomers and seniors rely less on online health information in order to manage their health because they would have less need? Third, will the presence of boomers and seniors that have more experience and familiarity with the Internet lead to greater use of online health information to manage health? Fourth, will individuals who are in a lower sociodemographic status rely less on online health information because of lack of resources to access this information? Finally, will avid Internet users use online health information more often to manage their health because they would have greater access to and familiarity with the Internet?

The American health care system is different from many Western countries, since it is administered primarily by the private marketplace. The majority of the United States population contracts with a private provider for his or her health insurance coverage. Medicare is a federal health insurance program for people age 65 and older. In addition, Medicaid, a program sponsored by the federal government and administered by states, is intended to provide health care and health-related services to low-income individuals. However, there are millions of Americans who do not fit into either the Medicare or Medicaid plans and, essentially, remain uninsured. Online health information is especially important, given the millions of uninsured Americans trying to get information on their health situation. Individuals can use this online health information to make informed choices on their health care needs. They potentially can use information on the Internet to better manage their health.

Essentially, has online health information influenced the behaviors of boomers and seniors with respect to their health care needs? This influence could be as extensive as visiting a doctor or simply talking to family or friends about health information that a boomer or senior found online.

Access to timely and reliable information on health and health care long has been a goal for seniors, who face a greater number of health conditions and use prescription drugs and health care services at a higher rate than younger adults (Kaiser Family Foundation, 2005). However, the online behavior of seniors has not been studied as closely as that of health information searches of adolescents (Gray, Klein, Noyce, Sesselberg, & Cantrill, 2005), women (Pandey, Hart, & Tiwary, 2003), cancer patients (Eysenbach, 2003; Ziebland, 2004), those affected by the digital divide (Skinner, Biscope, & Poland, 2003), and those that compare online and off-line behavior (Cotton & Gupta, 2004). There is little empirical research that examines whether online health searches affect the management of health (Lueg, Moore, & Warkentin, 2003; Nicholas, Huntington, Williams, & Blackburn, 2001), one of the two objectives of this study. This study measures whether Internet health information changed the self-reporting behavior of boomers and seniors and does not specifically address change in health outcomes.

There are two reasons why this study does a comparison of both boomers and seniors. First, baby boomers represent future seniors, and by examining this age group, this study can provide some indication about what the future holds for the Internet and health information. Second, both boomers and seniors are in the greatest need of health information, since they are more prone to have health problems than other age groups.

This study is different from existing works of Nicholas et al. (2001), Lueg et al. (2003), and Huntington et al. (2004), since it focuses on the use of online health information in the management of health. This study focuses especially on comparing two groups, boomers and seniors, while the existing empirical work examines the entire Internet population. This study is different from studies that conduct a meta-analysis, which combine published results from different sources (Eysenbach, 2003). This research performs a statistical analysis that leads to conclusions that are different from the original dataset (Kaiser
Family Foundation, 2005). The aim of this study is not just to learn about the differences between boomers and seniors and access to online health information; it is to discern the magnitude of differences between these groups and the impact of factors such as awareness and feelings on health management.

In order to accomplish the goal of examining online health information and the management of boomers’ and seniors’ health, this article is divided into several sections. First, this research examines the literature on the use of the Internet as a health information source. Second, this article outlines how the literature can be summarized into hypotheses that model the most probable impacts on management of boomers’ and seniors’ health. Third, this research provides details of the Kaiser Family Foundation’s e-Health and the Elderly dataset that is used to model public opinion data of online health information (Kaiser Family Foundation, 2005). The fourth and fifth sections discuss the models and results of tests on the use of online health information for health care management. The sixth section provides a discussion that outlines how the test results confirm or deny the specified hypotheses and shows the broader significance of this work. The last section provides avenues for future research and presents limitations of this study.

**LITERATURE REVIEW**

The following section outlines the common themes found in the literature on the Internet and health information and the management of health. They can be divided into the factors of differences between age groups, the health of the individual, online proficiency, sociodemographic characteristics, and awareness and feelings about online health information. Existing research shows that little is known about how Internet usage, health status, and sociodemographic characteristics affect health information seeking (Cotton & Gupta, 2004).

Eysenbach (2003) provides a conceptual framework of the possible link between Internet use and cancer. Some of the important factors, according to that author’s meta-analysis, indicate that Internet use is related to communication, community, and content, leading to an impact on cancer outcomes. In a similar line of inquiry, a study by Lueg et al. (2003) provided a conceptual framework of Internet searches for online health information. These authors examine the situations with which individuals find themselves confronted in terms of health needs and frequency of use, predicting access to health information. Eysenbach’s (2003) conceptual framework is different from Lueg et al. (2003), in that the former examines the social

---

Figure 1. Conceptual framework of access to online health information by boomers and seniors
aspects of Internet use and health information, while the latter study focuses on the situation involvement and the frequency of Internet use. The conceptual framework of this study is similar to Eysenbach (2003) and Lueg et al. (2003) but differs, in that it examines boomers and seniors and factors such as frequency and satisfaction having an influence on the management of health (Figure 1).

Online Health Information and Managing Boomers’ and Seniors’ Health

If the Internet can be used to change the behavior of individuals, this is one assessment of the long-term utility of this information resource. If individuals just look at information online and do not use it in any substantial way, it does not make much sense to invest in Internet health resources. The Internet suggests a remarkable change from the traditional doctor-knows-best approach (Eysenbach & Jadad, 2001). The Internet can be seen as challenging hierarchical models of information sharing, in which the provider of the information decides how the information should be delivered (Ziebland, 2004).

For example, existing research in a 2001 survey showed that 44% of online health information seekers said that the information they found online affected a decision about how to treat an illness or to cope with a medical condition (Fox & Rainie, 2002). A majority of respondents to a survey of adolescents and use of Internet health information reported that it helped them to start a conversation with a lay or professional medical person (Gray et al., 2005). Online health information seekers mostly are going online to look for specific answers to targeted questions (Fox & Rainie, 2000). In addition, four out of 10 young people say that they have changed their personal behaviors because of health information that they obtained online (Rideout, 2002). In an Internet survey, more than one-third of the respondents said that their conditions had improved after having visited a Web site, and more than one in four said that the Web information had resulted in a deferred visit or had actually replaced a visit to the doctor (Nicholas et al., 2001). Therefore, the existing research has examined adolescents’ and all age groups’ behavioral changes but has not focused on seniors and their use of the Internet for managing health. There are five factors outlined in the literature, which are differences between boomers and seniors, the health of boomers and seniors, their online proficiency, their sociodemographic characteristics, and awareness and feelings toward online health information that are predicted to have an impact on whether online health information is used.

Boomers’ and Seniors’ Differences

Existing empirical evidence shows that health seekers are proportionately more middle-aged than very young or old, with the highest proportion of usage witnessed in those between the ages of 30 and 64 (Fox & Rainie, 2000). A more recent survey indicates that 70% of baby boomers has gone online in 2004, while only 31% of seniors has gone online (Kaiser Family Foundation, 2005). Boomers will retire shortly, and the amount of online health information for which they will search should increase dramatically compared to what seniors are currently consuming. In order to explore both the present and what the future will hold for online health information and seniors, it is important to compare both age groups. In addition, individuals over the age of 50 may have a greater need for more information on health care than someone much younger because of the greater chance of facing health problems (Brodie et al., 2000).

Health of Boomers and Seniors

The literature also mentions that individuals who are in worse health will want to search more for
online health information. The Internet becomes an additional tool in order for them to search for health information. Empirical evidence shows that there is a link between an individual’s health and his or her need for online health information. Less healthy individuals are more likely to explore different aspects of a Web site and to use more health-related interactive features and, in doing so, improve their well being (Lueg et al., 2003). Individuals who were suffering with an illness were two and half times as likely, compared to respondents without a standing illness, to say that they had used information from the Internet as an alternative to seeing their general practitioner (Nicholas et al., 2001). In many cases, information seekers were acting on behalf of others, such as family and friends. However, access to online health information also should be related to the consumer’s ability to use the Internet, not just on whether they are healthy.

**Online Proficiency**

The ability to use the Internet also should have an impact on whether boomers and seniors use online health information to manage their health. Individuals who use Internet information more to manage their health have broadband Internet access, are frequently online, spend many hours online, and search for information on many different topics. Research shows that individuals using a Web site regularly were more likely to have said that the information was helpful (Nicholas et al., 2001). A survey of adolescents shows that there are issues of the disparity of Internet access and quality of Internet access such as dial-up vs. broadband connection (Skinner et al., 2003).

**Sociodemographic Characteristics of Online Health Seekers**

Another factor explored in the literature that should have an impact on access to online health information is the sociodemographic characteristics of the individual. There is research on the digital divide, between the haves and the have-nots of Internet access. This research predicts that those who have greater access to the Internet would have more resources in society. For instance, those groups of individuals who are more disadvantaged economically in the United States would have less access to the Internet and online health resources. Hispanics, the largest minority group in the United States, traditionally have had less Internet access (Fox & Rainie, 2000). Those with medium to high family incomes should be able to access the Internet more for health information because of greater resources.

Existing research shows that individuals who are older, have lower incomes, are minorities, are less educated, and are males will be less likely to use the Internet for health information seeking (Cotton & Gupta, 2004; Anderson, 2004). In contrast, women increasingly rely on the Internet to supplement health information received from traditional sources (Pandey et al., 2003) and are more likely than men to seek online health information (Fox & Rainie, 2000; Nicholas et al., 2001). Awareness and feelings toward online health information also should have an impact on using this information to manage consumers’ health.

**Awareness and Feelings Toward Online Health Resources**

A final factor that should explain access to Internet health information is the awareness and feelings of the individual toward online health information. If boomers and seniors have more positive feelings about the Internet as a health information resource, they will utilize it more often than someone who harbors more negative feelings toward the Internet. In addition, individuals who go online for health information frequently should use it more to manage their health. If a boomer’s or senior’s doctor or medical professional recommends or uses the Internet as a communication device,
the patient is more likely to use it to manage his or her health. In summary, the prediction is that boomers and seniors who are more aware of online health resources should use these resources more to manage their health. In addition, boomers and seniors who have positive feelings about the benefits of online health information will use this resource more to manage their health.

Empirical evidence shows that there is a relationship between using the Internet more often and accessing health information (Lueg et al., 2003). Those using the Web once a day were twice as likely to report that it helped a lot in terms of being better informed from health information found on the Web (Nicholas et al., 2001). E-mail is still a new medium for obtaining access to consumer health information and also is explored in the research as a way to manage a consumer’s health (Huntington et al., 2004). The literature just outlined can be formally specified with the following hypotheses that demonstrate the relationship between boomers and seniors, the Internet, and the management of health.

**HYPOTHESES**

In order to examine whether online health information has affected the choices that individuals make in managing their health and the differences between boomers and seniors, several hypotheses are tested in this article. These hypotheses are derived from the literature mentioned in the previous section and are divided into five areas:

**Boomers’ and Seniors’ Differences**

**Hypothesis 1:** Online health seekers who are baby boomers are more likely to believe that online health information has helped them to manage their health better compared with seniors.

**Health of Boomers and Seniors**

**Hypothesis 2:** Online health seekers who are healthy or who have family and friends that are healthy will rely less on online health information because of lack of need.

**Online Proficiency**

**Hypothesis 3:** Online health seekers who have broadband Internet access will go online more for health information to manage their health.

**Hypothesis 4:** Online health seekers who go online more often and conduct many online activities will use Internet health information more to manage their health.

**Sociodemographic Status**

**Hypothesis 5:** Boomers and seniors who are females will rely more on online health information.

**Hypothesis 6:** Boomers and seniors who are college educated will rely more on online health information.

**Hypothesis 7:** Boomers and seniors who are Hispanics will go online less for health information.

**Hypothesis 8:** Boomers and seniors who have family income above $75,000 will go online more for health information.

**Awareness and Feelings Toward Online Health Resources**

**Hypothesis 9:** Online health seekers who most of the time and always look to see who provides medical information on the Internet will use
online health information more to manage their health.

**Hypothesis 10:** Online health seekers who access health information online once or twice a month will have a greater likelihood of using online health information to manage their health.

**Hypothesis 11:** If a doctor has recommended a Web site to an online health seeker, he or she is more likely to use health information to manage his or her health.

**Hypothesis 12:** If an online health seeker has communicated with his or her doctor via e-mail, he or she is more likely to use online health information to manage his or her health.

**Hypothesis 13:** Online health seekers that have more positive feelings about looking for health information on the Internet are more likely to use this information to manage their health.

These hypotheses are examined with a dataset that surveyed public opinion of both baby boomers and seniors on their use and acceptance of online health information.

**DATASET AND METHODS**

The e-Health and the Elderly dataset is a nationally representative random digit dial telephone survey of 1,450 adults age 50 and older. Included in this sample were 583 respondents age 65 and older. The survey was designed by Kaiser Family Foundation (KFF) (2005) in consultation with Princeton Survey Research Associates (PSRA), and the survey was administered in the field by PSRA. The survey interviews were conducted between March 5 and April 8, 2004. The entire dataset of 1,450 respondents was first examined to determine the characteristics of boomers and seniors and access to online health information.

Out of the 1,450 responses to the survey, this study also has taken a subsample of 628 respondents, of which there were 464 boomers and 164 seniors surveyed. Therefore, the original dataset was split, and the sample sizes differ for both age groups. The 628 boomers and seniors represent those individuals who are called online health seekers. They both have Internet access and have looked for online health information. This group is of interest, since in this study, there is a comparison of the characteristics of those that actually look up online health information.

In this study, we use a consumer survey to explore the differences between boomers and seniors and their use of online health information to manage health. This research uses both descriptive statistics and logistic regression to explore differences in access to online health information between boomers and seniors.

**DESCRIPTIVE STATISTICS OF BOOMERS AND SENIORS AND ONLINE HEALTH INFORMATION**

In order to model the relationship between seniors and boomers, online health information, and its impact on managing health care needs, this study has specified the following variables that will comprise the models tested.

Table 1 provides information on boomers and seniors who go online for health information. Boomers that go online for health information represent 78.2%, while seniors that go online for health information represent just over 21% of those surveyed in this category. This table generally supports the notion that boomers tend to go online more for health information than seniors. Boomers that do not go online for health
information represent 43%, and seniors that do not go online represent 58% of those surveyed in this category.

Table 2 outlines demographic information of boomers and seniors that go online and do not go online for health information. The digital divide is very evident with the data presented in this table. For instance, 43% of college-educated individuals go online for health information compared with only 15% who are college-educated that do not go online for health information. Among females and Hispanics, there is not much of a difference in the percentage who go online and do not go online for health information. However, boomers and seniors that have a family income above $75,000 in 2003 were more likely to go online for health information. Finally, age seems to have an impact on accessing online health information. The mean age was 61 years for individuals that go online and 69 years for consumers who do not go online for health information. Higher income implies greater use of online health information.

### Table 1. Boomers and seniors who go online for health information

<table>
<thead>
<tr>
<th>Go online for health information (Yes or No)</th>
<th>Age group</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>50-64</td>
<td>335</td>
<td>42.5</td>
</tr>
<tr>
<td></td>
<td>65+</td>
<td>454</td>
<td>57.5</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>789</td>
<td>100</td>
</tr>
<tr>
<td>Yes</td>
<td>50-64</td>
<td>464</td>
<td>78.2</td>
</tr>
<tr>
<td></td>
<td>65+</td>
<td>129</td>
<td>21.8</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>593</td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 2. Demographic information of boomers and seniors and going online for health information

<table>
<thead>
<tr>
<th>Go online for health information (Yes or No)</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>College educated</td>
<td>822</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Gender is female</td>
<td>822</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Race is Hispanic</td>
<td>822</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Family income 2003 above $75,000</td>
<td>822</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>822</td>
<td>68.79</td>
</tr>
<tr>
<td>Yes</td>
<td>College educated</td>
<td>628</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Gender is female</td>
<td>628</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Race is Hispanic</td>
<td>628</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Family income 2003 above $75,000</td>
<td>628</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>628</td>
<td>60.88</td>
</tr>
</tbody>
</table>
The Internet, Health Information, and Managing Health

and having a college education means a greater likelihood of going online for health information. This finding also indicates that boomers are more likely to go online for health information, since the average age range was just over 61 years old.

Logistic regression is used to test whether sociodemographic variables predict whether boomers or seniors go online for health information. Logistic regression was used, since this study models dependent variables that are binary, represented by either a 1 or 0 (Nicholas et al., 2001; Lueg et al., 2003). The odds ratio can be used to interpret the relative impact of the observance of a 1 in the dependent variable. Table 3 shows that almost all of the sociodemographic variables help to explain whether someone goes online for health information, with the only exception being Hispanic. For instance, having a college education means that a boomer or senior is four times more likely to go online for health information. Having a higher income indicates that boomers and seniors are two times more likely to go online for health information. However, as the age of the respondent increases, this marginally decreases the likelihood of someone going online for health information.

### Dependent Variables

Table 4 provides a list of the dependent and predictor variables and also demonstrates whether there were differences between boomers and seniors in these variables. Perhaps the most important dependent variable is whether “somewhat” or “a lot” of information on the Internet has helped to take care of a senior’s or a boomer’s health. The mean score indicates that 59% of boomers and 46% of seniors believed that the Internet has helped them to take care of their health, demonstrating some impact on the management of their health.

The second dependent variable measures whether online health seekers had a conversation with family or friends about health information that they found online (Table 4). Family and friends who go online for health information may guide someone else as to whether they should see a doctor because of this information (Eysenbach, 2003). The results indicate that 66% of boomers said that they had a conversation with family members or friends, and only 48% of seniors said that they had this conversation about the information they saw online. There were statistically significant differences between seniors and boomers for

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Go Online for Health Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor Variables</td>
<td>Odds Ratio</td>
</tr>
<tr>
<td>Age</td>
<td>0.95</td>
</tr>
<tr>
<td>College educated</td>
<td>4.05</td>
</tr>
<tr>
<td>Gender is female</td>
<td>1.25</td>
</tr>
<tr>
<td>Race is Hispanic</td>
<td>1.06</td>
</tr>
<tr>
<td>Family income 2003 above $75,000</td>
<td>2.25</td>
</tr>
<tr>
<td>Constant</td>
<td>11.34</td>
</tr>
<tr>
<td>Nagelkerke R-Square</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Note: *** significant at the 0.01 level.
## Table 4. Difference of means tests of dependent and predictor variables for online health seekers; boomers are significantly different from seniors

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Mean of Boomers</th>
<th>Standard Deviations</th>
<th>Mean of Seniors</th>
<th>Standard Deviations</th>
<th>Probability Significantly Different Boomers and Seniors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somewhat and a lot of information on Internet helped take care health</td>
<td>0.59</td>
<td>0.49</td>
<td>0.46</td>
<td>0.50</td>
<td>0.07</td>
</tr>
<tr>
<td>Had a conversation family or friend about online health information</td>
<td>0.66</td>
<td>0.47</td>
<td>0.48</td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>Online health information changed behavior</td>
<td>0.36</td>
<td>0.48</td>
<td>0.25</td>
<td>0.43</td>
<td>0.00</td>
</tr>
<tr>
<td>Made a decision about how to treat an illness because of online health information</td>
<td>0.34</td>
<td>0.48</td>
<td>0.26</td>
<td>0.44</td>
<td>0.00</td>
</tr>
<tr>
<td>Visited a doctor because of information found online</td>
<td>0.16</td>
<td>0.37</td>
<td>0.13</td>
<td>0.34</td>
<td>0.07</td>
</tr>
<tr>
<td><strong>Predictor Variables: Health</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Excellent or very good health</td>
<td>0.58</td>
<td>0.49</td>
<td>0.52</td>
<td>0.50</td>
<td>0.07</td>
</tr>
<tr>
<td>Health problems index</td>
<td>3.95</td>
<td>2.45</td>
<td>3.80</td>
<td>2.41</td>
<td>0.48</td>
</tr>
<tr>
<td><strong>Predictor Variables: Online Proficiency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broadband Internet access</td>
<td>0.43</td>
<td>0.50</td>
<td>0.29</td>
<td>0.45</td>
<td>0.00</td>
</tr>
<tr>
<td>Online more than 10 hrs week</td>
<td>0.28</td>
<td>0.45</td>
<td>0.23</td>
<td>0.42</td>
<td>0.00</td>
</tr>
<tr>
<td>Online every day</td>
<td>0.57</td>
<td>0.50</td>
<td>0.51</td>
<td>0.50</td>
<td>0.08</td>
</tr>
<tr>
<td>Online activities index</td>
<td>2.86</td>
<td>0.97</td>
<td>2.45</td>
<td>1.14</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Predictor Variables: Online Health Information</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most of the time and always look to see who provides medical information on Internet</td>
<td>0.40</td>
<td>0.49</td>
<td>0.24</td>
<td>0.43</td>
<td>0.00</td>
</tr>
<tr>
<td>Access health information online once or twice a month or greater</td>
<td>0.38</td>
<td>0.49</td>
<td>0.34</td>
<td>0.48</td>
<td>0.04</td>
</tr>
<tr>
<td>Doctor recommended a health or medical Website</td>
<td>0.06</td>
<td>0.23</td>
<td>0.04</td>
<td>0.20</td>
<td>0.13</td>
</tr>
<tr>
<td>Communicated with doctor or other health care provider through email</td>
<td>0.12</td>
<td>0.32</td>
<td>0.11</td>
<td>0.31</td>
<td>0.55</td>
</tr>
<tr>
<td>Positive feelings about looking for health information on the Internet Index</td>
<td>2.59</td>
<td>0.73</td>
<td>2.23</td>
<td>0.94</td>
<td>0.00</td>
</tr>
<tr>
<td>Negative feelings about looking for health information on the Internet</td>
<td>0.68</td>
<td>0.78</td>
<td>0.71</td>
<td>0.81</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Notes: The number of observations are 464 for boomers and 164 for seniors.
this question with the reported F-statistic being significant at the 0.01 level, meaning that boomers were more likely to have a conversation with family and friends about health information that they found online.

The third dependent variable measures whether online health information changed the behavior of boomers and seniors (Table 4). Thirty-six percent of boomers’ behaviors changed as a result of online health information, compared with 25% of seniors. This result also was shown to have a statistically significant difference between boomers and seniors at the 0.01 level. Around one-third of boomers changed their behaviors, which is a good indication that the information that they are finding is affecting their health.

A fourth management of health issue was whether boomers or seniors made a decision on treatment of an illness as a result of the information they found online (Table 4). The results showed that 34% of boomers believed that they made a decision about how to treat an illness because of information they found online, while only 26% of seniors made a decision on treatment. This result also showed a statistically significant difference between the two age groups at the 0.01 level.

Another dependent variable was visiting a doctor as a result of the health information found online (Table 4). Only 16% of boomers visited a doctor as a result of health information they found online, while 13% of seniors visited a doctor. Visiting a doctor was the least utilized change in behavior as a result of online health information.

Referring back to Hypothesis 1 on whether online health information has been used to manage a boomer’s or a senior’s health, this study has found that overall, there were differences between both groups of online health seekers (Table 4). The mean values for all five dependent variables were higher for boomers compared with seniors. In addition, three out of the five dependent variables showed statistically significant differences between boomers and seniors at the 0.01 level. With these dependent variables outlined, this research also should describe the predictor variables and their characteristics.

**Predictor Variables**

The predictor variables used to explain how the Internet has managed the health care of boomers and seniors also are presented in Table 4. Many of the predictor variables are represented in terms of binary numbers in order to capture the specific impacts on the dependent variables. As previously noted, this study has divided the hypotheses into the differences between boomers and seniors, the relative health of the individual, his or her online proficiency, and how active he or she is at seeking online health information. This study discerns the impact that these factors have on the management of the health care of boomers and seniors.

To see all of the predictor variables, refer to Table 4. We will only mention a few of them in this section. For instance, an index was created of the health problems that boomers and seniors or someone they know have faced in the past year. An individual who has more health problems or is concerned with someone else’s health problems would score higher on the index. The health problems index indicates less than four issues that they or someone they know faced, indicated by online health seekers (out of nine possible health problems). The nine possible health problems listed were cancer, heart disease, obesity and weight loss, arthritis, diabetes, Alzheimer’s, high cholesterol, osteoporosis, and mental health.

The online activities index measures the amount of activities that boomers and seniors conduct online, and the average is around two activities (Table 4). The prediction is that health seekers who conduct more online activities have a greater likelihood of using health information to manage their health because of their familiarity and comfort with the Internet. The four
online activities that comprised the index were using instant messaging, reading news, buying a product, and checking the weather.

The online health information predictor variables also show the capacity of the individual to look up health information on the Internet (Table 4). Seniors are more trusting of the health information that they read online, with only 24% of seniors “most of the time” and “always” looking to see who provides medical information on the Internet. On the other hand, 40% of boomers are looking to see who provides the online health information. This difference was also statistically significant at the 0.01 level. In addition, boomers are more frequent consumers of online health information, using it at least once or twice a month, as represented by 38% of the sample. Seniors consume online health information marginally less frequently with 34% doing so.

With regard to overall positive feelings toward the Internet, there was an average score of two on an index scaled from zero to three, indicating that boomers and seniors have overall positive feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.

On an index of zero being the lowest and two being the highest, less than one was found, indicating that very few online health seekers harbor negative feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.

With regard to overall positive feelings toward the Internet, there was an average score of two on an index scaled from zero to three, indicating that boomers and seniors have overall positive feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.

On an index of zero being the lowest and two being the highest, less than one was found, indicating that very few online health seekers harbor negative feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.

With regard to overall positive feelings toward the Internet, there was an average score of two on an index scaled from zero to three, indicating that boomers and seniors have overall positive feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.

On an index of zero being the lowest and two being the highest, less than one was found, indicating that very few online health seekers harbor negative feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.

With regard to overall positive feelings toward the Internet, there was an average score of two on an index scaled from zero to three, indicating that boomers and seniors have overall positive feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.

On an index of zero being the lowest and two being the highest, less than one was found, indicating that very few online health seekers harbor negative feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.

With regard to overall positive feelings toward the Internet, there was an average score of two on an index scaled from zero to three, indicating that boomers and seniors have overall positive feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.

On an index of zero being the lowest and two being the highest, less than one was found, indicating that very few online health seekers harbor negative feelings toward the Internet as a source of health information. The index was calculated by adding up the specific responses to whether the online health seeker agreed that online health information gave them information quickly, whether it helped them feel more informed when they go to the doctor, and whether it allows them to get information from a lot of different sources.


### Table 5. Logistic regression of factors predicting whether online health information has managed a boomer's or a senior's health

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Somewhat and a lot of information on Internet helped take care health</th>
<th>Had a conversation family or friend about online health information</th>
<th>Online health information changed behavior</th>
<th>Made a decision about how to treat an illness because of online health information</th>
<th>Visited a doctor because of information found online</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boomer = 1 (age between 50 to 64)</td>
<td>1.31 (4.49)**</td>
<td>1.49 (10.67)**</td>
<td>1.71 (14.99)**</td>
<td>1.66 (12.63)**</td>
<td>0.90 (0.36)</td>
</tr>
<tr>
<td>Health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Excellent or very good health</td>
<td>1.11 (0.88)</td>
<td>1.64 (21.39)**</td>
<td>1.21 (3.06)</td>
<td>1.36 (7.09)**</td>
<td>1.10 (0.41)</td>
</tr>
<tr>
<td>Health problems index</td>
<td>0.99 (0.36)</td>
<td>1.14 (34.51)**</td>
<td>1.04 (3.40)</td>
<td>1.06 (0.00)</td>
<td>1.04 (1.35)</td>
</tr>
<tr>
<td>Online Proficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broadband Internet access</td>
<td>0.82 (2.79)</td>
<td>0.63 (15.51)**</td>
<td>0.90 (0.80)</td>
<td>0.68 (9.37)**</td>
<td>0.88 (0.51)</td>
</tr>
<tr>
<td>Online every day</td>
<td>1.55 (13.13)**</td>
<td>1.31 (5.01)**</td>
<td>1.82 (22.22)**</td>
<td>1.43 (7.15)**</td>
<td>1.41 (3.99)**</td>
</tr>
<tr>
<td>Online more than 10 hrs week</td>
<td>0.65 (9.62)**</td>
<td>0.89 (0.71)</td>
<td>0.97 (0.05)</td>
<td>1.46 (7.26)**</td>
<td>0.81 (1.39)</td>
</tr>
<tr>
<td>Online activities index</td>
<td>1.01 (0.01)</td>
<td>0.93 (1.50)</td>
<td>0.95 (0.55)</td>
<td>0.92 (1.83)</td>
<td>0.80 (7.75)**</td>
</tr>
<tr>
<td>Online Health Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access health information online once or twice a month or greater</td>
<td>3.51 (115.95)**</td>
<td>1.59 (17.35)**</td>
<td>2.29 (54.99)**</td>
<td>3.11 (95.83)**</td>
<td>1.82 (16.15)**</td>
</tr>
<tr>
<td>Communicated with doctor or other health care provider through email</td>
<td>0.91 (0.28)</td>
<td>1.95 (12.45)**</td>
<td>1.58 (7.28)**</td>
<td>2.83 (35.37)**</td>
<td>1.86 (9.94)**</td>
</tr>
<tr>
<td>Doctor recommended a health or medical Website</td>
<td>0.76 (0.97)</td>
<td>0.91 (0.10)</td>
<td>1.17 (0.31)</td>
<td>2.13 (7.10)**</td>
<td>3.68 (22.13)**</td>
</tr>
<tr>
<td>Positive feelings about looking for health information on the Internet</td>
<td>2.46 (119.44)**</td>
<td>1.63 (46.32)**</td>
<td>1.85 (42.97)**</td>
<td>1.76 (31.68)**</td>
<td>1.84 (19.73)**</td>
</tr>
<tr>
<td>Most of the time and always look to see who provides medical information on Internet</td>
<td>1.31 (4.86)**</td>
<td>2.80 (63.13)**</td>
<td>1.61 (15.80)**</td>
<td>1.57 (13.23)**</td>
<td>1.66 (11.21)**</td>
</tr>
<tr>
<td>Negative feelings about looking for health information on the Internet</td>
<td>0.75 (16.82)**</td>
<td>0.81 (10.16)**</td>
<td>0.71 (21.06)**</td>
<td>0.70 (20.83)**</td>
<td>0.77 (6.21)**</td>
</tr>
<tr>
<td>Constant</td>
<td>0.08 (89.93)**</td>
<td>0.13 (65.14)**</td>
<td>0.03 (122.43)**</td>
<td>0.04 (103.03)**</td>
<td>0.03 (66.25)**</td>
</tr>
<tr>
<td>Nagelkerke R-Square</td>
<td>0.27</td>
<td>0.22</td>
<td>0.22</td>
<td>0.26</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Notes: ** significant at the 0.05 level and *** significant at the 0.01 level.
is not that high, given the attention placed on the differences between these age groups (Kaiser Family Foundation, 2005).

In terms of online proficiency, those who had broadband Internet access were less likely to have a conversation with a family member or a friend about what they saw online and less likely to make a decision about how to treat an illness. However, for those online health seekers who are online every day, this had a consistent impact across all five dependent variables, if they used online health information to manage their health. For instance, there was a 1.82 times greater chance that daily online users changed their behavior because of online health information. In addition, individuals who go online every day were 1.41 times more likely to visit a doctor because of information that they found online. In terms of being online more than 10 hours a week, this had a negative likelihood for the dependent variables “somewhat” or “a lot” of information on the Internet helped to take care of their health, but had a positive impact with an odds ratio of 1.46 for making a decision on how to treat an illness. Overall, there was no overwhelming support that being more proficient with the Internet helped to take care of their health, but had a positive impact with an odds ratio of 1.46 for making a decision on how to treat an illness. Overall, there was no overwhelming support that being more proficient with the Internet helped to take care of their health, but had a positive impact with an odds ratio of 1.46 for making a decision on how to treat an illness.

The strongest predictors of using health information to manage health were for the online health information awareness and feelings variables. Frequent consumers of online health information were 3.51 times more likely to use the Internet to take care of their health. It has become a valuable tool for their health care management needs. There was also a 3.11 greater likelihood of someone making a decision about treating an illness to use the Internet more than once a month for health information. In fact, frequently accessing health information registered an impact for all of the dependent variables.

Individuals who had positive feelings about online health information were more likely to use online health information to manage their health. Boomers and seniors who had more negative feelings toward online health information were less likely to use it for health management. This finding was consistently found across all five of the dependent variables. Online health seekers who usually looked to see who provided the health information were more likely to use this to manage their health. For instance, individuals who looked to see who provided the health information were 2.80 times more likely to have a conversation with family or friends about the information that they saw online. If a doctor recommended a health or medical Web site, online health seekers were 2.13 times more likely to make a decision about how to treat an illness because of online health information, and they were 3.68 times more likely to visit a doctor because of information they found online. In addition, individuals who communicated with a doctor or a health care provider via e-mail were 1.95 times more likely to have a conversation with a family member or a friend about health information that they found online. Overall, the logistic regression results indicate the most consistent and highest support for
increased awareness and positive feelings toward online health information as a driver for helping to manage boomers’ and seniors’ health.

CLASSIFICATION TREES ANALYSIS OF TAKING CARE OF BOOMERS’ AND SENIORS’ HEALTH

Another way to examine the relationship between boomers and seniors and online health information is a classification tree analysis. Classification trees are used to predict membership of cases or objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables. Classification tree analysis is a common data mining technique. Figure 2 shows that the taking care of health variable is related to boomers and seniors having positive feelings toward looking for online health information. In addition, having positive feelings about online health information is related to accessing health information online once or twice a month or more. These findings reinforce the logistic regression results that awareness and feeling toward online health information helps in the management of a boomer’s or a senior’s health. How do these findings relate to the hypotheses outlined in the beginning of this article?

DISCUSSION OF RESULTS AND HYPOTHESES

This section will discuss how the empirical results of this study confirm or deny the hypotheses (see Table 6). First, the evidence shows through the difference of means tests that boomers and seniors are different in terms of their use of online health information in the management of their health. The mean values scored higher for boomers in using online health information to take care of their health, having a conversation with a family or a friend about the health information found online, changing their behavior because of online health information, and making a decision about treating an illness because of online health information. There is some evidence that boomers will use more online health information to manage their health, supporting Hypothesis 1. However, there is no overwhelming support for differences between boomers and seniors and using health information to manage their health, with boomers only utilizing this information one and one-half times more than seniors.

The health of the individual, or Hypotheses 2, only predicted the use of online health information to manage a boomer’s and a senior’s health when they had a conversation with a family member or a friend about online health information and when they made a decision about how to treat an illness. In addition, the health variable predicted a boomer’s and a senior’s behavior when he or she talked to his or her doctor about information that he or she found online. Overall, there was not overwhelming support that the health of the boomer or senior had an impact on use of online health information to manage his or her health.

Hypotheses 3 and 4 examine whether being more proficient online means that the online health seeker will use the Internet more to manage his or her health. The results consistently showed that those who go online every day were more likely to use the Internet to manage their health. There was not much support that being more proficient with the Internet meant that boomers and seniors would use it more often for health information, since many of the other independent variables in this category were not statistically significant. This is most likely explained by what these online health seekers are doing on the Internet; they are looking for information, which does not require, for instance, broadband Internet access, since a standard dial-up connection will suffice. Therefore, this research cannot confirm that being more proficient with the Internet means that
Figure 2. Classification tree of impact of online health information taking care of a boomer’s and a senior’s health
The Internet, Health Information, and Managing Health

Table 6. Support for hypotheses of boomers and seniors and online health information

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>Supported? (Yes, No, or Partially)</th>
<th>Major Test(s) Performed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis 1: Online health seekers who are baby boomers are more likely to believe that online health information has helped them manage their health better compared with seniors.</td>
<td>Yes</td>
<td>Descriptive statistics, difference of means tests, and logistic regression</td>
</tr>
<tr>
<td>Hypothesis 2: Online health seekers who are healthy or who have family and friends that are healthy will rely less on online health information because of lack of need.</td>
<td>Partially</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 3: Online health seekers who have broadband Internet access will go online more for health information to manage their health</td>
<td>No</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 4: Online health seekers who go online more often and conduct many online activities will use Internet health information more to manage their health.</td>
<td>Yes</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 5: Boomers and seniors who are females will rely more on online health information.</td>
<td>Yes</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 6: Boomers and seniors who are college educated will rely more on online health information.</td>
<td>Yes</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 7: Boomers and seniors who are Hispanics will go online less for health information.</td>
<td>No</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 8: Boomers and seniors who have family income above $75,000 will go online more for health information.</td>
<td>Yes</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 9: Online health seekers who most of the time and always look to see who provides medical information on the Internet would use online health information more to manage their health.</td>
<td>Yes</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 10: Online health seekers who access health information online once or twice a month would have a greater likelihood of using online health information to manage their health.</td>
<td>Yes</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 11: If a doctor has recommended a Website to an online health seeker he or she is more likely to use health information to manage their health.</td>
<td>Partially</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 12: If an online health seeker has communicated with his or her doctor via email he or she is more likely to use online health information to manage their health.</td>
<td>Yes</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Hypothesis 13: Online health seekers that have more positive feelings about looking for health information on the Internet are more likely to use this information to manage their health.</td>
<td>Yes</td>
<td>Logistic regression</td>
</tr>
</tbody>
</table>
online health seekers will use this communication media to manage their health more than those who are not as proficient. However, this could change with greater availability of streaming video health information, which is much more suited to a broadband Internet connection.

Boomers and seniors of higher sociodemographic status use the Internet for health information much more than lower sociodemographic status individuals (Hypotheses 5-8). Therefore, this research confirms that there is a digital divide in access to online health information, and public policy should attempt to address this issue. It should be noted that one-third of the United States adult population has not gone online and, therefore, would not be able to benefit from online health information (Pew Internet & American Life, 2005).

Hypothesis 10 was confirmed in the logistic regression results that those who are frequent patrons of online health information actually will use it more often to manage their health. The results showed that health seekers who are accessing health information online once or twice a month or more will be more likely to actually use this information to manage their health. Therefore, these individuals are not just searching for information; they actually are using some of what they find online. Individuals who have positive feelings about online health information also will use it more often to manage their health, and individuals that harbor more negative feelings will use it less often (Hypothesis 13). Online health seekers who are very aware of who provides the medical information on the Internet are more likely to use this information to manage their health (Hypothesis 9). Finally, where there is communication with their doctor via e-mail or at the doctor’s office about online health information, boomers and seniors are more likely to use online health information to manage their health (Hypothesis 12). Awareness and feelings toward online health information generally were well-supported predictor variables (Hypotheses 9-13), having an impact on a boomer’s or a senior’s use of this information to manage his or her health care needs.

RECOMMENDATIONS, LIMITATIONS, AND FUTURE RESEARCH

This article examined the use of the Internet for accessing health information by boomers (age 50 to 64) and seniors (age 65 and over). Boomers generally use Internet health information to manage their health more than seniors. However, there was no overwhelming differences between boomers and seniors, which is the main difference in the finding from another study (Kaiser Family Foundation, 2005). For instance, boomers are much more likely to talk to a doctor about health information that they saw online. Boomers are around one and one-half times more likely than seniors to use online health information to manage their health. This study found that awareness and feelings toward online health information provided the best explanation of health information for management of boomers’ and seniors’ health.

Since boomers were found to use online health information marginally more than seniors, what are the implications of this observation? Will seniors of tomorrow be similar to seniors of today? Perhaps boomers will continue to seek online health information as they get older. The implication that boomers and seniors may be in the greatest need of health information may not be true in the future with the growing obesity epidemic in the United States, which affects all age groups.

Some policy recommendations should be noted in order to bring more seniors online and to enhance the quality of Internet health resources. Health care professionals should recommend Web sites, promote more effective search and evaluation techniques, and be more involved in developing and promoting uniform standards for
The Internet, Health Information, and Managing Health

health Web sites (Morahan-Martin, 2004). Since only a minority of seniors has ever gone online, this represents a significant digital divide. These findings confirm that for the foreseeable future, the Internet is less likely to be a primary source of information for most seniors, which suggests a need to invest more heavily in education and outreach strategies. This is especially the case for seniors with low or modest incomes, who are least likely to go online for this information. These recommendations could make seniors more aware and could create a positive experience when going online for health information.

In the near future, the Internet will become a decision-making tool for seniors, who will need to make choices about the Medicare prescription drug benefits. They will need to decide which plan has the most attractive premium and to determine whether it will cover the medications they take and will work with the pharmacy they use. Seniors also will need to manage the Internet to make these important decisions. Web site design is part of the solution, since seniors have problems scrolling on Web sites and remembering Web pages (Voelker, 2005).

There are some limitations of this research. With any type of public opinion data, especially when asking subjective questions about sensitive topics of consumers’ health, respondents may not be as forthcoming with information. Another limitation is that of the general applicability of the results, given that the proportion of the sample is different for seniors and boomers. In addition, there is no question that specifically addressed whether there was an improvement in the health outcome, just that people felt better informed. Future research could do a longitudinal follow up of this dataset, which might reveal shifts in the use of Internet health information for managing health with boomers and seniors, looking at other measures to see if there is an impact on change in the person’s health.

REFERENCES


**ENDNOTES**

1 In this study, for simplicity, baby boomers are classified as those individuals between the ages of 50 and 64, and seniors are classified as 65 and older.

2 This author would like to thank Victoria Rideout, M.A., Vice President and Director, Program for the Study of Entertainment Media and Health, Kaiser Family Foundation (KFF), for the dataset and documentation used in the statistical analysis of this study. I would also like to thank Virginia Rodgers for her editorial assistance.

3 For the data analysis, the software package used was SPSS version 13.0.

*This work was previously published in the International Journal of Healthcare Information Systems and Informatics, edited by J. Tan, Volume 1, Issue 2, pp. 20-38, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 4.20
The Online Effect: Transitioning from the Legacy Help Desk to the Online Task Management System

Kym Mawson-Lee
University of New South Wales, Australia

ABSTRACT

This case centres on a request management system that was developed and implemented in response to the growing requests to the Information Technology (IT) unit for assistance. Primarily, this assistance was with a core corporate application called WorkDesk, a back office processing system. The request management system, an IT unit initiative, was designed in collaboration with representatives of the organisation’s business units. The initial development project lasted for approximately two months and upon implementation was well received by the organisation. This first implementation turned out to be a proof of concept for a much larger system called the Request System that has been developed and implemented as a single online channel to manage all requests made of the IT unit. The Request System has greatly increased the ability of the IT unit to strategically manage IT and its support across the organisation. The core benefit of this case is the discussion of the successful development of an online task management system and its 10 key functions. This system effectively manages all requests made to the IT unit and introduced accountability for its completion. The system described can be easily repurposed based upon the knowledge demonstrated in this case study for a wide range of task management purposes. The situation described in this case resulted in the organisation’s maturing in terms of IT understanding, an issue with which many organisations are struggling and one that, if understood, will assist students to have a positive impact on similar organisations.

ORGANISATIONAL BACKGROUND

The Office of the Employment Advocate (OEA) was established under the Australian Workplace Relations Act (1996). The Employment Advocate’s role includes providing assistance and advice to employees and employers on workplace legisla-
tion, especially related to Australian workplace agreements (AWAs) and freedom of association, particularly in small businesses (OEA, 2003). The OEA, an agency of the Department of Employment and Workplace Relations (DEWR), is based in Sydney with regional offices in each state and the Northern Territory and consists of 180 staff. For a recent 2004 summary of the OEA, see http://www.oea.gov.au/graphics.asp?showdoc=/corporate/newsletter-sept2004.asp&Page=0 (Accessed 11/26/04).

The OEA’s mission is improving Australian workplaces (OEA 2004). The primary way in which the OEA achieves this mission is through AWAs. AWAs are individual agreements struck between an employer and an employee about the employee’s wages and conditions of employment. AWAs were introduced by the Australian government to give employers and employees flexibility in setting wages and conditions, allowing them to customise arrangements to their workplaces.

In terms of management structure, the OEA is very hierarchical both physically and culturally for an organisation of 180 staff with two layers of management that report to two layers of executive. In order to manage the entire organisation, the OEA has implemented the organisational structure depicted in Figure 1. From this organisational chart, it is possible to see that the IT unit is located at the third level of the hierarchy. The IT unit is a subunit of Corporate, which is an internal organisational enabler considered a supporting business unit of the organisation. The purpose of Corporate is to assist the OEA

---

**Figure 1. OEA organisational chart (OEA, 2004)**

![Diagram](image-url)
The Online Effect

in meeting its business objectives as an internal enabler; the IT unit is considered one component of that. Having this structure has required the National IT Manager to carefully manage the Information Technology Steering Committee (ITSC) as the main channel of communicating with the rest of the executive level of the OEA. The ITSC comprises the National IT Manager (Chairperson), together with the two deputy Employment Advocates, the Corporate Director, and the Director — the Strategic Advice Unit. Outside of the ITSC, high-level strategic actions regarding IT are approved in consultation with the Corporate Director, an otherwise non-IT-related role. The IT unit is based in Sydney as part of the OEA head office and supports regional offices in each state and the Northern Territory supporting a total of 180 staff.

Financially, the OEA is funded via an Australian government budget allocation provided to the OEA through the parent organisation, DEWR. For budgeting purposes, internally all business units are operated as cost centres. The IT unit budget, in particular, is split into two core categories: infrastructure service from DEWR and normal operating costs. The OEA has grown steadily over the last seven years from approximately 30 staff to 180 (a 600% increase with 40% of this growth being in the last two years) this year with 30 roles being filled just in the last 12 months (OEA, 2003). A strategic plan is developed each year in order to meet organisational objectives. The OEA’s strategic plan is a three-year rolling plan, meaning it is updated on a yearly basis to address the next three years.

For further information about the OEA, please refer to www.oea.gov.au.

This case study is about a request system developed by the OEA. Primarily, this case will examine task management and coordination of these tasks within an IT environment, as well as the capture of the information related to these tasks. As an example, these requests currently are broken up into six key areas. The first area is for requesting a modification to a Web page either on the OEA intranet or the OEA Internet site. These requests can be as simple as changing a word to something as complex as developing a new section on the intranet for a new team and all of the related pages that go with it. The next four areas are for requests specific to four different corporate applications that the IT unit support for the OEA, and the final area is for suggestions relating to anything that the IT unit can implement to improve the organisation.

The OEA and the IT unit are approximately seven years old, and until 12 months ago, the IT unit handled most requests as ad hoc issues with no formalised process for handling them. The IT unit was staffed with six individuals whose average tenure was three years, with most of the knowledge being held tacitly (i.e., knowledge that is not explicit and highly personal). Approximately 12 months prior to the writing of this case, two of these key staff left the agency, taking with them the knowledge of how to provide key IT functions. The IT unit supports four core corporate applications as well as an intranet and Internet site. This case describes one of the ways that the IT unit planned to manage this transition in staff.

SETTING THE STAGE

Within the OEA, there are varying levels of IT utilisation from a cutting edge .NET Internet application to an archaic Visual Basic (VB) application. This VB application contains vital organisational data that only a few people can access and that was found to be impossible to migrate to a more accessible system. The management of IT within the agency has a hierarchical approach. Representatives from the many levels attend a monthly ITSC meeting where a range of IT-related strategic issues are discussed, such as new projects and their related business cases. The National IT Manager chairs these discussions and prepares the agenda and minutes. Decisions
are made at the ITSC level and are carried out primarily by IT staff. The other members of the ITSC have varying levels of IT knowledge.

In the past, as an organisation, the OEA has not placed a high priority on IT. Previous experiences of IT projects had left the executive and management teams with a poor view of IT and its utilisation. One example of this is WorkDesk, the OEA’s core corporate application. WorkDesk is both a workflow and content management system that has tailor-made user interfaces, functions, and workflow specifically designed for the purpose of filing, assessing, and approving AWAs. More than half of the organisation’s staff is required to use WorkDesk on a daily basis. This application cost the equivalent of the annual IT budget to build and implement but was found wanting in some of its basic functions, such as searching and retrieving some specific types of AWAs within complex queue arrangements, causing many users to contact the IT unit on a daily basis for assistance to find these AWAs. Another example is the client service system (CSS), which cost approximately 5% of the annual IT budget and was developed in-house. This application is a basic customer management system used by the Client Services Network (CSN) business unit. During its informal development, it suffered massive scope creep, and while the application essentially is successful, some staff felt that the system should have been developed following a more formal process.

This was a difficult situation for the IT unit, as it did not have any allocated budget for the development of new systems; however, as the following scenario will show, there was a definite need.

Twelve months ago (2003), the IT unit of the OEA underwent a major change in staff. One of the key concerns in this situation was the loss of corporate systems support knowledge. After this change of staff occurred, a diverse range of skills was available; however, four of the staff were relatively new and inexperienced. The departure of two key staff at a time when there were few established or documented processes and only moderate application documentation created operational challenges. When the primary WorkDesk administrator left the IT unit, it became clear that there was little understanding about what support the previous administrator was providing. This required a carefully managed transition to the new WorkDesk staff. WorkDesk being an idiosyncratic application required a vast amount of knowledge to be maintained effectively.

In order to manage this transition effectively, a consolidated approach to handling requests was required. The aim was for the IT unit to be able to service the OEA’s business units quickly, effectively, and professionally. It was from these requirements that the Request System was born. With the implementation of this system, the IT unit was able to move from a reactive position to a position of effectively managing their workload. This system is the subject of this case study and is called the Request System.

The main issue confronting the IT unit at this time of staff turnover was that it required an effective way to manage requests for support. The IT unit now services a massive variety of requests, ranging from help-desk requests for WorkDesk and expert system design style requests to requests for designing whole applications such as the CSS. The ongoing management of this variety of requests was quite complex. The Request System was proposed to handle these requests and overcome this complexity.

Initially, the Request System was required to handle the capture, notification, processing, and completion of requests for assistance with WorkDesk. However, very quickly, the system was enhanced to facilitate requests for the following:

- Web modifications to both the Internet and intranet (Webmods);
- AWAonline II, an Internet application;
- AWAMS 1, a legacy application;
- Other organisational applications: COMS, CLOCS, CSS, System Monitor; and
Suggestions relating to AWAonline II, intranet, Internet, online polls, and the ITU Request System itself.

For some months prior to the Request System, requests would come to IT staff in all manner of ways, including e-mail, telephone, hallway chatter, chance meetings, during meetings, mobile phone, and third-hand messages. The before and after of the methods requests for assistance were handled as presented in the diagram presented in Figure 2.

Some of the challenges associated with the Before diagram (Figure 2) and having so many avenues to make requests include:

- No record of the request.
- No centralised archive of those requests that were recoded.
- No way to track the amount of workload currently being experienced by IT staff.
- Little opportunity for management reporting.
- No process documentation; therefore, little opportunity for process improvement.
- No knowledge retention and management.

In the After diagram, the communication flows are marked by the letters $a \rightarrow e$, which signify the task artefact. As an example of the flow of a task artefact, each request for assistance with AWA-
online II is assigned to a staff member, who has the necessary skills to complete a request about AWAonline II. Importantly, once the request is completed, the original requester is notified of its completion.

In summary, the development of the Request System started in May 2003. The impetus for the project was to better manage IT staff turnover and to overcome the ad hoc manner by which requests for assistance were being made to the IT unit, combined with the ad hoc manner by which these requests were handled. There was little reassurance from the staff involved as to whether a task had been completed or whether the business unit was happy with the actions taken to complete the task.

**CASE DESCRIPTION**

In early 2003, the OEA IT unit had found itself in a challenging situation. There was no funding for systems enhancement or development such as the Request System, and staff with important tacit knowledge about handling requests for core corporate applications were about to leave the organisation. The members of the IT unit with support of the National IT Manager worked hard to overcome these challenges. As a starting point, the project manager of the request system development needed to identify what functions were needed and if such a system for managing these task-based requests existed. Two key literature sources — collaborative systems and knowledge management literature — were used, and an action research methodology was adopted. The goal was to identify a way to assist the OEA to facilitate the management of tasks being requested of IT staff. This case will follow the development and implementation of this application, which has centred on the development of an effective e-based task artefact (an electronic instance and record of a task).

Supporting an organisation’s teams in routine and ad hoc tasks is a continual goal of Information System (IS) (Mahling & Craven, 1995). However, for every IS-related solution developed and implemented, there are generally some drawbacks. Influenced by the role of technology in business processes, this case begins to develop a continuous thread, drawing upon collaborative systems theory for current and future e-based information systems development. In particular, the case study reviews the support for facilitating, requesting, documentation, communication, and overall management of tasks.

Considerable literature utilises the concepts of artefact and task artefact; in particular, the fields of Computer Supported Cooperative Work (CSCW), Interorganisational Information Systems (IIS), and Software Engineering (Daneshgar, 2001; Dustdar, 2004; Giaglis, Paul & Doukidis, 1996; Schulz & Milosevic, 2000). However, it appears that very little has been done to combine these disparate uses of the concept into a single holistic thread in order to establish task artefact utilisation as part of task management, a fundamental research component of IS. It is intended that this case will begin to address this issue and form the basis of future work. From initial readings, there seemed to be no unified approach to e-based task management within the literature.

E-based support for task management dates back to the origins of computing machinery (Mahling & Craven 1995) and provides a solid foundation for this case study. As e-business becomes adopted more widely, one of the emergent research areas is the identification of models that are generalizable in order to support the process of e-business (Mukhopadhyay, Smith, & Muniz, 2003). The particular aspect of e-business support that this case covers is that of task management within e-based task management systems in order to support requests for IT support.

This case also begins to evaluate the usefulness of designing a model for e-based task artefacts
in order to facilitate and improve the process of task management within the OEA. Furthermore, in order to support the integration of task management within business processes, e-business task management systems require clarity in the identification of task artefacts and how these artefacts can be used to support OEA processes. As an example, a fairly simple task artefact is provided in Figure 3.

In response to needs of the system requirements identified so far, an action research approach was taken to facilitate the iterative development of the request management system. The action research methodology adopted was based on McKay

Figure 3. Example of a task artefact generated from the WorkDesk request form

<table>
<thead>
<tr>
<th>Task Artefact</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>REQUEST DETAILS</td>
<td>Note: A screen shot of the artefact interface is provided in Appendix 1a and 1b.</td>
</tr>
<tr>
<td>Request Number : WKD3118</td>
<td>&lt;- unique for each request occurrence</td>
</tr>
<tr>
<td>Open Date : 11/05/2004 2:38:49 PM</td>
<td>&lt;- To assist tracking and reporting</td>
</tr>
<tr>
<td>Request Assigned : BOB</td>
<td>&lt;- Role or specific staff member</td>
</tr>
<tr>
<td>Contact Number : (02) 9246 0587</td>
<td>&lt;- Requester phone number</td>
</tr>
<tr>
<td>Error Type : WorkDesk Application</td>
<td>&lt;- category of issue</td>
</tr>
<tr>
<td>Folder Reference Number or Part 1 Number : 180794</td>
<td>&lt;- unique reference information</td>
</tr>
<tr>
<td>EXTRA INFO : I am unable to log into workdesk. When I left my PC I was logged in, and now that I have returned, I do not have an icon on my toolbar, and when I try to log in there is a msg reading “This user is already logged on.”</td>
<td></td>
</tr>
</tbody>
</table>
and Marshall (2001) with specific attention to their linked spiral model that develops research knowledge based upon the actions being taken. These authors provide an excellent illustrative example of this process, of which a representation is provided in Figure 4. For further information on this model and how it is implemented, please refer to McKay and Marshall (2001).

The Request System needed to handle requests for any type of IT assistance as well as the consolidation and recording of these requests.

![Image](image_url)
A core requirement of the system was the capturing of all of the required information that the IT unit needed in order to service a request within what is called a task artefact. Initially, the scope of the application was to manage and store requests from business units in relation to one specific corporate application, WorkDesk. The scope for the project quickly grew to encompass requests in relation to all corporate applications as well as the OEA intranet and Internet sites. The road map in Figure 5 provides an overall view of the path taken — Stages 1, 2, and 3 correspond to the investigation and analysis sections of the Systems Development Lifecycle (SDLC) and stage 4 to the design and implementation sections. More detail is provided in the Development Process section that follows.

**Development Process**

The OEA already had a corporate intranet. In order to facilitate ease of access for the entire organisation, the request system was to be built into the intranet. The intranet was based on Active Server Page (ASP) technology, which seemed fairly ideal as a development environment, particularly since it was a known technology. The following three phases are the iterative stages in the development of the Request System.

**Phase 1**

While conceptually each of the IT staff had their own idea of what should be developed, it was important to develop a shared understanding...
among IT staff and across the organisational business units. The project team facilitated the following steps:

1. The IT staff work-shopped the following:
   a. Requirements in terms of what information needed to be captured in order to complete a request for assistance for the WorkDesk application; and
   b. Some indicative categories into which requests for assistance with WorkDesk could be broken down.

2. With the guidance of the project team, expert users of WorkDesk work-shopped the following:
   a. The types of questions that they ask regularly;
   b. Those questions that tend to be novel; and
   c. The overall requirements of such a system.

On the basis of these workshops, the first component of the system was developed. It consisted of a limited interface with a corresponding backend database. This component, which also served as the basic prototype for what was to become the Request System, was tested by IT staff and two designated expert users for a period of two weeks. During this time, all information about each request was collected, not only in the task artefact via the Request System but also comments made about the request and additional e-mails that were sent. After two weeks, the collected data were analysed with the goal of facilitating 100% of all requests for WorkDesk assistance through the request system. A series of recommendations was made on how to improve this component, which was then prioritised and developed. This improved component was then made available to those staff that had contributed in the first instance to the requirements specification for further comment.

This is a very brief description of many hours and weeks of work that resulted in the first version of the WorkDesk component of the request system. As this was the first attempt to consolidate the channels for making requests within the organisation, it was important that a clearly defined model for developing the interfaces for each of the other OEA applications be developed in order to make subsequent development easier. The knowledge developed from this process is presented at the end of phase three as well as how other organisations could utilise these aspects far more easily. The finalised interface for inputting the required information for the WorkDesk task artefact is presented in Appendix 1a.

Phase 2

Based upon the knowledge from phase one, phase two commenced soon after the completion of version one of the first component. Phase two was the development of a new component that would facilitate requests for changes to be made to the OEA Internet and intranet sites known as Webmods. This phase was made somewhat easier by the existence of a previous database that was used by the Webmod staff to help them keep track of the requests made. However, this database was not accessible to non-IT staff and essentially involved a lot of copying and pasting from e-mails received. The project team replicated the process from phase 1, identifying the requirements of the staff who carry out the Web modifications and then working with regular requesters of Web modifications to identify their requirements.

From the requesters’ points of view, their requirements were simple: (a) make the requested change to the Web site and (b) let them know when it was completed. The main challenge was making the job of updating Web pages easier and more efficient. The Webmod staff requirements included which page needed changing, whether there were other pages affected by this change,
and when the changes needed to occur? This essentially meant asking the requester to supply considerably more information with the request than normally was provided in the first instance. Once this range of requirements was identified, version one of the Webmod request component was developed. The finalised interface for inputting the required information for the Webmod task artefact is presented in Appendix 1b.

Phase 3

The third phase was to roll out the request system to all OEA employees. The system to this point contained a component for WorkDesk requests and for Webmod requests. A third feature was added to drive further developments of the system; namely, Other Requests. The motivation behind this was to ensure that all requests made to the IT unit could be received via one channel — the request system. It was the role of the project team to process these Other Requests by breaking them down into existing or new request categories and having requirements specified for the development of an appropriate new task artefact interface, if required.

During this third phase, collaborative systems literature, particularly that of Dustdar (2004) and Dustdar and Gall (2003), was employed to take full advantage of previous research work in the area. As this literature was reviewed, key characteristics were identified and recorded into a task artefact characteristics chart. This chart is presented in Appendix 2 to assist readers in identifying the basis of what was built into each task artefact and the system overall. Each of the characteristics was assessed for suitability for inclusion in the request system. The main criterion by which each of the characteristics was assessed was whether its adoption would assist users or administrators to make or complete requests more effectively. Interestingly, many of the characteristics identified in the literature had already been developed into the request system intuitively as part of the requirements identified by members of the IT unit. However, there were also quite a few characteristics, such as notification and status, that enabled a more effective end-to-end process to be established.

Once these three phases were completed, the Request System project team began the process again, although not so formally and not in as much depth. The purpose of the iterations was to calibrate the effectiveness of the system, which is summarised in the following 10-point summary of the key aspects of the Request System. The iteration of all three phases occurred three to four times in varying levels of depth.

The task artefacts provide the basis of the success of the system and much of the knowledge gained. These task artefacts started as a blank file asking high-level questions. As a result of the three-phase iterative action component of the action research approach adopted, the task artefacts now had the following core functionality and contributed to the research component. These ideas can be draw upon for other task support systems:

10 Core Functions for E-Based Task Artefacts

1. Each task artefact is prepopulated with identifying information of the task requester, such as name, contact number, time stamp of the request, IP address, and physical address.
2. The requests are categorised on the basis of the required outcome to facilitate relevant detailed questions being presented automatically to the person requesting the task. This is done in order to capture all of the pertinent information required for completing each task.
3. The requester can set a date by which the request needs to be completed, which is
particularly useful for work prioritisation. A review date is assigned automatically to all Webmods so that all OEA intranet and Internet Web pages will be reviewed on a six-month basis.

4. Prior to submitting selected requests, the requester is presented with a Frequently Asked Questions (FAQ) page that includes explanations of how requesters can complete a common range of requests without IT staff intervention, if, for example, a user receives an unexpected response from an application.

5. The request is given a unique identifier. This request ID is sent immediately to the requester with a copy of the request just made.

6. Each request is assigned automatically to an IT staff member on the basis of the type of request made and the skills it will require to be completed. This ensures that the staff member with the required skill will complete the request. This is managed by a skills matrix, identifying all IT staff and their associated skills. Each IT staff member is able to navigate via the intranet to a page identifying all of the outstanding and completed requests.

7. The staff member who has been assigned the request then can set a status to the request such as analysing, or if more information is required, there is an option to ask for further specific information and then set the request to User Reply Required, at which point the original requester is notified by e-mail that further information is required.

8. Requesters are able to view the status of their requests at any time through a Web interface.

9. When requests are completed, there is an e-mail notification to the requester, advising the requester that the IT unit believes the request is completed, but that if the requester believes it is not completed, the requester is asked to reply to the e-mail with further details of the request.

10. Each of the requests is counted by category and by whether it is completed or not. These counts are fed into a reporting component that is used to assess the IT unit’s performance on a monthly and quarterly basis, facilitating a more strategic approach to IT management.

In terms of the research component of action research, most of the work to date has been related to identifying literature that is of assistance and developing the task artefact characteristics chart provided in Appendix 2.

**Summary of System Outcomes**

The three-phase approach to the development of the Request System has proved to be effective. This effectiveness is demonstrated through the greatly reduced number of follow-up phone calls and e-mails that previously were required to complete the root cause analysis often required to complete tasks. The Request System also assisted the IT unit in identifying opportunities for innovative ways to support the other OEA business units by facilitating the analysis of request patterns and request types. If there is a significant increase in a particular type of request, this now can be managed in a more effective manner. Often, any request, even those that were trivial to complete, required a phone call or an e-mail to identify specific information; however, now, only highly complex requests require further information, which is facilitated by a requester-reply-required status that automatically sends an e-mail to the original requester for further information.

From the reporting module of the Request System, for the last six months, the system has been handling between 300 and 400 requests per month. The system has approximately 40 regular users (more than one request per week) and approximately 100 less regular users. Importantly,
prior to the Request System, it was unknown how many requests for assistance IT staff was handling each month. An unplanned benefit is that the IT unit managers now have concrete evidence of what work the IT unit staff is doing and can ensure consistent responses to requests made. For illustrative purposes, a screen shot of the July 2004 report is available in Appendix 1c.

Key aspects of the request system’s success came from two core business decisions. One consisted of ensuring that the application was designed to effectively handle processes from end to end. This facilitated the second decision: making the system usage mandatory. The decision for mandatory usage created a lot of buy-in from the business units, as their representatives had a vested interest in ensuring that they understood and communicated the needs of their business unit in terms of making requests. The project team was able to make it clear that those requests that had not been clearly specified in terms of the business requirements would not be handled by the request system and, therefore, could not be requested. When combined with the commitment of the developers, this led to the development of a system that has received excellent feedback from users both in terms of ease of use and completeness of how the tasks were handled.

As a result of the Request System, the following three goals were achieved:

1. **Centralised location for self-help information.** The Request System has provided the ideal location for answers to commonly asked questions
2. **Reporting.** Real-time reporting of all requests is made available to all staff. The report provides details ranging from an individual request to a report on the number of requests received in relation to a specific issue.
3. **Correct assignment of requests.** Previously, various requests would be made to staff members who did not have the knowledge to complete the requests. Now, the request is assigned automatically to the correct IT staff member on the basis of skill sets that are contained in the skills matrix component of the system.

### CURRENT CHALLENGES/PROBLEMS FACING THE ORGANISATION

The current stage of the Request System is that of maturing. Business units are starting to leverage the concepts utilised in the system for their own benefit, creating task artefacts for their own use and for requesting tasks of business units other than IT. However, some staff members within the organisation are still only beginning to familiarise themselves with the opportunities that the Request System provides.

Within the organisation, the IT unit holds regular meetings with other business units in order to explore and address issues that arise as a result of requests made in the Request System. Some of the issues experienced so far include the following:

- Some users do not utilise the request system as designed. They use incorrect categories and do not supply the information specified in the user interface.
- Users sometimes do not read information readily available in the system to resolve issues. For example, those requests that are covered in the FAQ would allow users to resolve their own request.
- Some users lodge requests that they could have completed themselves, because they prefer IT staff to complete the task to save their own time.
- Users lodging requests that are not IT tasks. IT staff has to forward such a request to another business unit for completion.
The Online Effect

- Non-IT staff members who have requests assigned to them do not have their performance measured by completion of such requests. As a result, some of the non-IT staff who are trialing the use of the system for their own purposes have many outstanding requests, which reduces the overall perception of the effectiveness of the system.

- There are significant differences in the requests from experienced users and those received from new users of the request system. As an example, there are varying levels of clarity in the request description. This may cause misunderstandings between the IT staff and the requester.

- Despite best efforts in communicating the existence and usage of the system, some staff members are still not sure of its use and tend to use other methods to resolve their requests.

- There is room for more executive involvement in terms of reporting.

While this list provides an understanding of the day-to-day issues facing the OEA, at a strategic level, there is a more complex issue. The Request System has been somewhat evolutionary in its design; as staff members within the organisation become aware of its effectiveness and add their own new task artefacts online, it has become somewhat confusing for some staff to use. The IT unit developed a plan to tackle this by involving the business units in developing a new menu and navigation structure; however, as is often the case in IT development, this has been challenging, since these business units have other tasks that they consider to be a higher priority. Another related issue is the user interface. Due to original design restrictions, the layout and colors are quite bland; however, the original restrictions no longer apply, and a new layout has been designed. The IT unit is ready to release this new design but can’t until it addresses the menu structure.

Key benefits for students and managers include a description of iterative application development using action research, using business requirements driving application development, and drawing upon collaborative systems research when designing a task management system.

REFERENCES


APPENDIXES

1a. Workdesk task artefact interface — Note those fields that are currently populated are always populated with the requesters’ information gleaned from their PC login credentials

1b. Webmod task artefact interface — Note those fields that are currently populated are always populated with the requesters’ information gleaned from their PC login credentials
The Online Effect

1c. Screen shots of the reporting module
### Task artefact characteristics chart

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Details</th>
<th>Example</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRUD</td>
<td>Create, Read, Update and Delete.</td>
<td></td>
<td>Lindsay (1997)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Hyperlinks</td>
<td>Linking all artefacts and processes (people and tasks). This is the simplest method for implementing links, they are distributed with each document, performs well in terms of scalability and disconnected operation.</td>
<td><a href="http://www%E2%80%A6">http://www…</a></td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Versioning</td>
<td>Facilitating visibility of artefacts over time. Accurate planning and identification of existing resources often require the continued visibility of older versions of artefacts.</td>
<td>1.2.3 current design release 3. pre-release testing increments</td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td>Notification</td>
<td>When the resource changes Completion of an activity.</td>
<td>send email as notification</td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td>Unique ID</td>
<td>Differential between similar instances.</td>
<td>AWA12</td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td>Accessibility</td>
<td>Visibility of an artefact with a unique identifier should depend on both the role permissions of the artefact and the permissions of the user.</td>
<td>Approved or denied</td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td>Linking of artefacts</td>
<td>When retrieving an artefact, other related artefacts are identified and presented</td>
<td>All artefacts created by bob.</td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td>Reusability</td>
<td>Facilitate re-use and transmission of organisational knowledge between engineers working with the artefact and consequently greater productivity through collaborative learning</td>
<td>Identify repeated issues.</td>
<td>Boldyreff et al. (2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Boldyreff et al. (2002)</td>
</tr>
<tr>
<td>Identification of any one performing a creation, update or a deletion.</td>
<td>This will ensure a task can be completed for someone.</td>
<td>John Smith</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Domain/platform/storage mechanism independent</td>
<td>Adopt commonly used technology</td>
<td>Web and html</td>
<td>Li et al. (1994); Dustdar (2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Recordability</td>
<td>This will enable reusability</td>
<td>Record in a database</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Auditable</td>
<td>This will enable a process to be tracked, which in turn will enable improvement</td>
<td>Record a range of characteristics</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Timestamp</td>
<td>At the time of creation, to enable auditing</td>
<td>Time of creation hh:mm:ss</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Routing of artefacts</td>
<td>Enable processes to be established</td>
<td>Who task was assigned to and when</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dustdar (2004)</td>
</tr>
</tbody>
</table>
### 2b. Task artefact characteristics chart (continued)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Details</th>
<th>Example</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity recording</td>
<td>Recording the people who create, read, update and complete the task artefact and the role they occupy.</td>
<td>John Smith</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Access rights and rules</td>
<td>Defining who can create, read update and complete</td>
<td>Jill can view bobs requests</td>
<td>Dustdar (2004); Lindsay (1997)</td>
</tr>
<tr>
<td>Process tracking and progress</td>
<td>What path did the task artefact follow to be completed?</td>
<td>A status characteristic would achieve this.</td>
<td>Dustdar (2004); Lindsay 1997</td>
</tr>
<tr>
<td>Reporting</td>
<td>Not discussed but will be investigate further as part of this research.</td>
<td>How many tasks requested today? = 4</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Retrieve share and distribute.</td>
<td>This is about accessibility and distribution in order to complete a task. This will enable completion of a task and reporting.</td>
<td>Send all type a tasks to bob.</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Traceability of associated activities, work items, performed actions</td>
<td>To be explored</td>
<td>Who accessed task AWA999</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Changes regarding states of activities are visible</td>
<td>Tracking of the task state.</td>
<td>Celia completed task 8</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Security management</td>
<td>(Authorisation and access control),</td>
<td>Bob can see all task requests</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Mail and notification management,</td>
<td>Notification of new task request.</td>
<td>Sally requests a task and bob gets an email notification</td>
<td>Dustdar (2004)</td>
</tr>
<tr>
<td>Manage and define access rights to all</td>
<td>Caramba objects: Read (R), Insert (I), Update (U), and Delete (D); (2) definition of security roles (not to be confused with the Roles used in Organizational Objects) and the association between security roles and rights and (3) definition of the association between access rights and accounts.</td>
<td>Authorization is granted by a login which is encrypted, where the account login name and password have to be provided</td>
<td>Lindsay (1997); Dustdar (2004)</td>
</tr>
<tr>
<td>Template</td>
<td>Support the capture of required information. Ease and consistency</td>
<td>Task requirements template</td>
<td>Kennedy and Schauder (1998)</td>
</tr>
<tr>
<td>Status</td>
<td>Describing what state is the task artefact in.</td>
<td>“analysing”</td>
<td>Heuvel and Van den Maamar (2003); Lindsay (1997)</td>
</tr>
</tbody>
</table>
### 2b. Task artefact characteristics chart (continued)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Details</th>
<th>Example</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keepers</td>
<td>Functionalities of storage and access to shared data (persistency).</td>
<td>Save, access rights, accessibility.</td>
<td>Ellis and Maltzahn (1997)</td>
</tr>
<tr>
<td>Communicators</td>
<td>functionality related to explicit communications among team members</td>
<td>Messaging systems (e-mail).</td>
<td>Ellis and Maltzahn (1997)</td>
</tr>
<tr>
<td>Coordinators</td>
<td>ordering and synchronization of individual tasks that make up a whole</td>
<td>Workflow Management Systems</td>
<td>Ellis and Maltzahn (1997)</td>
</tr>
<tr>
<td>Team-Agents</td>
<td>(semi)intelligent software components that perform domain specific</td>
<td>Email filters</td>
<td>Ellis and Maltzahn (1997)</td>
</tr>
<tr>
<td>Metadata</td>
<td>Intrinsically part of the artefact, Ideally automatically populated</td>
<td>Time&lt;2:15&gt;</td>
<td>Boldyreff et al. (2002)</td>
</tr>
<tr>
<td>Re-use</td>
<td>Those with the correct access can view information within a task artefact</td>
<td>Check ho to complete a task that was previously completed</td>
<td>Boldyreff et al. (2002)</td>
</tr>
<tr>
<td>Transmission</td>
<td>Mostly of organisational knowledge between roles working with the</td>
<td>Send BOB the completed task 8.</td>
<td>Boldyreff et al (2002)</td>
</tr>
<tr>
<td>Inheritance</td>
<td>Artefacts be specialised from an abstract base type artefact</td>
<td>Fix network card is inherited from fix server</td>
<td>Boldyreff et al. (2002)</td>
</tr>
<tr>
<td>Searching</td>
<td>To find desired artefacts</td>
<td>Find all tasks on fixing servers</td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td>linking</td>
<td>To find related artefacts</td>
<td>All artefacts relating to a specific task</td>
<td>Fielding et al. (1998)</td>
</tr>
<tr>
<td>Exception Handling</td>
<td>The artefact must not cause error if there is an exception, or a new</td>
<td>No task being specified.</td>
<td>Mahling and Craven (1995)</td>
</tr>
<tr>
<td>Facilitate initial</td>
<td>Facilitate shared understanding by transferring current understanding and</td>
<td>Publication task, what is to be published</td>
<td>Mahling and Craven (1995)</td>
</tr>
<tr>
<td>Knowledge of ability</td>
<td>Knowledge of who can complete the task</td>
<td>Who can publish</td>
<td>Mahling and Craven (1995)</td>
</tr>
<tr>
<td>Automate Routine</td>
<td>Automatic capture of information already identifiable by the system</td>
<td>Time</td>
<td>Mahling and Craven, (1995)</td>
</tr>
<tr>
<td>Supporting Status</td>
<td>Display and allow search by status</td>
<td>Search by “completed”</td>
<td>Mahling and Craven Lindsay (1997)</td>
</tr>
<tr>
<td>System interrogation</td>
<td>Searching</td>
<td>Find all tasks that are not complete</td>
<td>Mahling and Craven (1995)</td>
</tr>
</tbody>
</table>
### Task artefact characteristics chart (continued)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Details</th>
<th>Example</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Granularity</td>
<td>Capture all required information regardless of the task</td>
<td>Exact url of an item to be published</td>
<td>Mahling and Craven (1995)</td>
</tr>
<tr>
<td>Robust language</td>
<td>All users need to understand the language the system capture for the artefact</td>
<td>Web Mod v Internet web modification</td>
<td>Mahling and Craven (1995)</td>
</tr>
<tr>
<td>Assign a person to the task</td>
<td></td>
<td>Bob must do task 333</td>
<td>Mahling and Craven (1995)</td>
</tr>
<tr>
<td>Design and use scenarios are conceptually different</td>
<td>Design and actual use of a system are never quite identical.</td>
<td>Once the e-based task artefact is implemented with all characteristics allow them to remain unused</td>
<td>Kutscha et al. (1997)</td>
</tr>
</tbody>
</table>

*This work was previously published in the Journal of Cases on Information Technology, edited by M. Khosrow-Pour, Volume 8, Issue 1, pp. 79-96, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
This section includes a wide range of research pertaining to the social and organizational impact of human-computer interaction around the world. Chapters introducing this section analyze gender differences in the adoption and reaction to IT, while later contributions offer an extensive analysis of differing social and cultural dimensions of technology adoption. The inquiries and methods presented in this section offer insight into the implications of human-computer interaction at both a personal and organizational level, while also emphasizing potential areas of study within the discipline.
Chapter 5.1
Gender, Race, Social Class and Information Technology

Myungsook Klassen
California Lutheran University, USA

Russell Stockard
California Lutheran University, USA

INTRODUCTION

The issue of the underrepresentation of women in the information technology workforce has been the subject of a number of studies and the gender gap was an issue when the digital divide dominated discourse about women’s and minority groups’ use of the Internet. However, a broader view is needed. That perspective would include the relation of women and IT in the communities in which they live as well as the larger society. The information society that has emerged includes the United States and the globalized economy of which it is an integral part.

Women and minorities such as African Americans and Latinos are underrepresented in computer science (CS) and other information technology positions in the United States. In addition, while they are no longer numerically underrepresented in access to computers and the Internet – as of 2000, (Gorski, 2001) - they continue to enjoy fewer benefits available through the medium than white boys and men. The following article explores the diversity within women from the perspectives of race, ethnicity and social class in North America, mainly United States.

The technology gender and racial gap persists in education and in the IT workforce. A broader and deeper look at women’s position in relation to the increasingly techno-centric society reveals that women may have reached equality in access, but not equity in academic study and job opportunities.

BACKGROUND

Linebarger (2003) pointed out three traditional digital divide constructs: ‘family socioeconomic status’, ‘location of access to new technologies’ and ‘gender/race’ for school age children. Inequalities tend to appear along both social class and...
gender lines, with male students and students from high socioeconomic status backgrounds well positioned to outpace female students and students from lower socioeconomic backgrounds in terms of computer skills and knowledge (Lockard, Abrams, & Many, 1987).

**Equality in Access, but No Equity in IT Jobs**

The gender digital divide refers to the gap in access rates between men and women (Gorsky, 2001). Based on this traditional gender digital divide definition, the gender digital divide gap has narrowed to reach “access equality.” In 2002, 83 percent of American family households owned a computer (Corporation for Public Broadcasting, 2004). About the same proportion of adult men and women had access to home computers. The digital connectedness of American families was increased through home computer ownership. In 2001, 59 percent of American people had connections at home. By the end of 2000, women surpassed men to become a majority of the United States online population (Gorski, 2001).

The societal race and gender gaps in the United States as a whole have narrowed in the past 10 years, but in the IT field, the gender gap generally appears to be wider at all levels of employment. Overall growth in these IT occupations was so strong during the decade of the 1990’s that women working in IT continued to increase through the year 1996. According to D’Agostino (2003), in 1996 women were 41 percent in the IT field. The ITAA (2003) recorded a decline to 34.9 percent by 2002.

The situation is worse in highly professional positions such as computer programmers and computer systems analysts, where women tend to lag far behind men. The Table 1 shows how women are overrepresented in lower IT positions while there are few women in professional computer science fields.

However, the potential exists for this situation to change. Kvasny (2003) reports that minority women in low-income communities perceive IT as a means of escaping poverty while highly educated, middle-class and professional women regard IT as offering fewer opportunities for advancement. Kvasny suggests that IT and gender studies recognize the diversity within women.

**Gains in Access, but Loss in Computer Science Major**

The problem of underrepresentation of women in IT starts from the math and science pipeline at school. Through high school, girls are less

<table>
<thead>
<tr>
<th>Information Technology Occupations</th>
<th>% Men</th>
<th>% Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer systems analysis and scientists</td>
<td>72.2 %</td>
<td>27.8 %</td>
</tr>
<tr>
<td>Operations and systems researchers and analysts</td>
<td>51.3 %</td>
<td>48.7 %</td>
</tr>
<tr>
<td>Computer programmers</td>
<td>74.4 %</td>
<td>25.6 %</td>
</tr>
<tr>
<td>Computer operators</td>
<td>53.2 %</td>
<td>46.8 %</td>
</tr>
<tr>
<td>Data entry keyers</td>
<td>18.3 %</td>
<td>81.8 %</td>
</tr>
<tr>
<td>Total IT occupations</td>
<td>65 %</td>
<td>35 %</td>
</tr>
</tbody>
</table>

(Source: Bureau of Labor Statistics)
Gender, Race, Social Class and Information Technology

likely than boys to enroll in computer science classes, and the disparity increases in programming courses.

The American Association of University Women (AAUW) commissioned early studies on the gender gap in education (1992, 1998). The first study noted the barriers faced by children from lower socioeconomic status. It also pointed out African American girls had fewer interactions with teachers even though they tried to initiate such interactions. By the latter AAUW study, the issue of technology had emerged. The report noted that a gender gap had begun to appear in computer science classes. Girls made up only a small proportion of students in such classes and the gap widen between grades eight and grades eleven. The study reported that boys exhibited a higher degree of self-confidence about computer skills than girls.

According to the National Council for Research on Women (2002), by the eighth grade, Latinas score higher in math than their male peers; and by twelfth grade they do better in science than Latinos, but they are outperformed by their male peers on the Math SATs.

A more alarming situation is the trend of fewer women entering the field of computer science. Between 1985 and 2002, women went from earning 36 percent of the computer science bachelor’s degrees (D’Agostino, 2003) to only 20 percent in 2002 (Taulbee, 2004). Even when women choose computer science as their major, their relative (compared to men) lack of preparation for the coursework and male dominated classroom climate forces them to drop out of the program (Margolis, 2003).

Table 3 shows the distribution of bachelor’s degrees in computer science by the race/ethnicity of the recipients for 2001. Analysis of the figures from a gender perspective reveals some interesting patterns. One is that the gender gap is greater among white women than women from underrepresented minorities. Generally, African American and Native American women and

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>White, non-Hispanic</td>
<td>18,479 (78%)</td>
<td>5,296 (22%)</td>
</tr>
<tr>
<td>Underrepresented Minorities</td>
<td>3,892 (59%)</td>
<td>2,663 (41%)</td>
</tr>
<tr>
<td>Black, non-Hispanic</td>
<td>2,182 (53%)</td>
<td>1,906 (47%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1,519 (69%)</td>
<td>680 (31%)</td>
</tr>
<tr>
<td>American Indian or Alaskan Native</td>
<td>191 (71%)</td>
<td>77 (29%)</td>
</tr>
<tr>
<td>U.S. Citizens and Permanent Residents</td>
<td>1,492 (73%)</td>
<td>549 (27%)</td>
</tr>
<tr>
<td>Unknown Race/Ethnicity</td>
<td>1,492 (73%)</td>
<td>549 (27%)</td>
</tr>
<tr>
<td>U.S. Citizens and Permanent Residents, Total</td>
<td>28,013 (73%)</td>
<td>10,517 (27%)</td>
</tr>
</tbody>
</table>

Table 2. Computer science degrees by gender

<table>
<thead>
<tr>
<th></th>
<th>Bachelor’s</th>
<th>Master’s</th>
<th>PhDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>80.6%</td>
<td>73.6%</td>
<td>83.2%</td>
</tr>
<tr>
<td>Female</td>
<td>19.4%</td>
<td>26.4%</td>
<td>16.8%</td>
</tr>
</tbody>
</table>

Table 3. Bachelor’s degrees in computer science degrees by race/ethnicity and sex of recipients. 2001 (Source: Taulbee, 2003)
Latinas earn more than their share of science and bachelor’s degrees than men in their respective groups (National Council on Research for Women, 2002).

White women earned 22 percent of the computer science bachelor’s degrees conferred on white males and females, while underrepresented minority women earned 41 percent of the CS bachelor’s degrees conferred on minority males and females. Among African American CS bachelor’s degree recipients, the numerical gap is almost non-existent, less than 300 out of a total of just under 4100. The gap between Latino/Hispanic males and female bachelor’s recipients is much larger than that among African Americans, 839 out of a total of 2,199. The picture is similar for American Indian CS bachelor’s recipients, with a 29 percent female to 71 percent male ratio.

Why don’t girls’ recent gains in access to technology translation into long-term advancement in college majors and careers? Gurur and Camp (2002) noted that attitudes, computer experience, computer games, mentoring and role models, self-confidence, computing environments, societal influence, teacher and family encouragement, all-female environments, perceived difficulties in balancing work and family are some factors contributing to the problem. The nerd image of the field, the fact that computer games are targeted mostly at boys, the perception of computing careers as boring, and the lack of role models for girls are a few reasons pointed out by Margolis and Fisher (2000).

Cohoon (2001) points out that despite beliefs espoused by some in computer science that women have deeply ingrained traits that suit them less for study and practice of computer science than men, female underrepresentation in computer science could be avoided. Gilbert has observed the importance of support systems for immigrant groups studying computer science and suggested that underrepresented groups in the U.S. could benefit from such practices (Loftus, 2004).

**Technology Access and Social Class/SES**

Here technology access will be discussed from two perspectives. One is computer access and the other is Internet access. Table 4 shows computer usage by 5-17 year-olds, either at home or at school by their parents’ education and by income (NCES, 2003). A large gap exists between children whose parents have the least and the most education. Rates of students using computers at school are fairly even among students from high-income families and those from low-income families, each 75.2 percent and 85.4 percent, respectively. Schools have played a key role in access equality for students of different backgrounds.

According to the National Center for Education Statistics (NCES, 2001), in 1994, 20 percent of

<table>
<thead>
<tr>
<th>Table 4. Computer usage, Ages 5-17, 2001 Source: NCES, 2003</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Income</strong></td>
</tr>
<tr>
<td>average</td>
</tr>
<tr>
<td>&gt;$75,000</td>
</tr>
<tr>
<td>&lt;$20,000</td>
</tr>
<tr>
<td><strong>Parent’s education</strong></td>
</tr>
<tr>
<td>Didn’t finish high school</td>
</tr>
<tr>
<td>Has post graduate degree</td>
</tr>
</tbody>
</table>
public schools in low income areas enjoyed Internet access while 35 percent of all public schools had such access. By 2000, the figures were 94 percent and 98 percent, respectively.

Of those who use the Internet, 78 percent access it from home, compared to 68 percent from school. Thirty-five percent of students from families with income of less than $20,000 access the Internet at home, while 52 percent do so at school (NCES, 2001). Again, schools have played an important role in narrowing digital divide by providing equal access to students from different class backgrounds.

When girls are compared to boys in terms of computer access, Internet access, the time spent on computers and Internet and computer activities, they are not much different (National School Boards Foundation, 2000). In fact, Linebarger and Chernin (2003) show that among children between 4 and 8 years old, boys use computers more than girls, but use Internet less than girls do.

What affects girls the most with respect to the digital divide is their SES, similar to the case for boys. There is research on correlations between gender and social class or SES and computer access. Davies, Hancock and Condon (2003) reported that there is little difference in access to home computers and access among families with high SES for boys and girls and but there is about a 10 percentage difference in home computer and Internet access between boys and girls in lower SES families.

Usage of home computers varies by social class or SES. Fifty percent of the children from high-SES families with home computers used word processing, compared with only 24 percent of the children from low SES families with computers at home (NCES, 2001).

Parents’ SES influences students’ attitudes toward computers and eventually contribute to widen the technological gender gap. Shashaani (1994) reported that family SES affected sex differences in attitudes towards computers. In general, gender-differential attitudes are more pronounced in the lower socioeconomic group and SES has a stronger effect on girls than boys. Both girls and boys perceive the gender stereotypes about computing held by their parents, and such attitudes inversely affect girls’ own attitudes.

**FUTURE TRENDS**

From a societal perspective, while girls’ and women’s use of computers and the is on a par with use by boys, it is yet to be seen if the comparable patterns of usage in the future will translate into equitable academic and professional outcomes. The computing environment, starting from the home through elementary, middle, and high schools must be made more encouraging (Cohoon, 2001 & Margolis, 2000). Further research on family SES with gender needs to be done. In addition race and gender interactions have not yet been a central theme among researchers, but research sponsored by the National Science Foundation is altering the research priorities. The findings by researchers such as Kvasny (2003) that working class African American women see opportunities in computer-related or IT careers in contrast to middle class white women’s perceptions of obstacles in the compute science pipeline need to be explored in further research.

**CONCLUSION**

While school provides good access to all students and computer access is theoretically equitable between the sexes, complex social factors serve to limit girls’ participation both in school and at home. Technology is more than physical resources; it is intertwined with social and cultural factors that differentially affect interactions. These include family computer cultures and encouragement, psychological access, social identities and the setting for the technology use, to name a few.
The traditional definition of “digital divide” no longer paints an accurate portrait of technology. The whole picture needs to be evaluated from a broader perspective. As computer prices continue to fall, lower SES families or schools in low SES areas may have greatly narrowed the traditional digital divide, but further gains may be difficult.

Simple class and economic predictions cannot be made, even within the white American segments. The process becomes much more complex as race and ethnic characteristics are included.

**Key Terms**

**Digital Divide:** the condition of unequal access to computer-related resources, varying along the demographic dimensions of age, gender, race and ethnicity, education, income, and nationality.

**Diversity:** in a social and cultural context, the presence in a population of a wide variety of cultures, opinions, ethnic groups, socioeconomic backgrounds, disabilities, and sexual preferences.

**Information Technology (IT):** a term that encompasses all forms of technology used to create, store, exchange, and use information in its various forms. Graduates of computer science majors may seek IT jobs.

**Math/Science Pipeline:** a phenomenon where the number of female students, students with lower socioeconomic status, and students of color in proportion to white males in advanced math and science progressively declines during high school.

**Social Class:** a category of people who have generally similar educational histories, job opportunities, and social standing and who are conscious of their membership in a social group that is ranked in relation to others and is replicated over generations.

**Socioeconomic Status:** the economic, social and physical environments in which individuals live and work, as well as demographic and genetic factors. Measures for SES may include: Income or Income Adequacy, Education, Occupation, or Employment.

**Underrepresentation:** a situation in which members of a group by color, race, sex, or ethnicity (but not all) constitute a lower percentage of the total number of people within the category in the job market, or in schools.

**References**


Gender, Race, Social Class and Information Technology


Gilbert, J. (2004). Building the Future Black Faculty Pipeline. Presentation at the @froGEEKS conference. Santa Barbara.


This work was previously published in Encyclopedia of Gender and Information Technology, edited by E. Trauth, pp. 705-710, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 5.2

Gender and the Culture of Computing in Applied IT Education

Susan C. Herring
Indiana University, USA

Christine Ogan
Indiana University, USA

Manju Ahuja
Indiana University, USA

Jean C. Robinson
Indiana University, USA

INTRODUCTION

The “shrinking pipeline” of women who ascend through the ranks in computer science education programs and careers is by now a familiar problem. Women drop out at rates faster than men at all levels of educational and professional advancement, resulting in a gender gap especially pronounced at the highest levels of the computing workforce, and that has not narrowed appreciably at any level in more than 20 years (Camp, 1997; ITAA, 2005; Vegso, 2005). Efforts to move more women into the pipeline at lower levels have met with limited success (cf. the Carnegie Mellon experience as reported by Margolis & Fisher, 2002); girls and women still express less interest than boys and men in studying computer science and pursuing information technology (IT) careers (Bentson, 2000; Vegso, 2005).

A reason often cited in the literature is the masculine culture of many computer science programs and IT workplaces, which is perceived by many women as alien and unwelcoming (Bentson, 2000; Spertus, 1991; Turkle, 1988). Even when institutions make efforts to treat women and men equally or accord women special consideration...
in admissions and hiring decisions, attitudes discouraging women from entering computing persist, both within the institutions and in society at large. Sometimes these attitudes are expressed overtly: Underground “hacker” culture is notoriously antagonistic to women (Gilboa, 1996), and even mainstream computer aficionados respond with resistance and sexist jokes to proposals to recruit more girls and women to study computer science (Slashdot.org, 2005). Moreover, there is a widespread perception that computer experts are socially-isolated “geeks” or “nerds” obsessed with technology, a mode of being that women, who tend to be more socially oriented, find unappealing (Margolis & Fisher, 2002; Turkle, 1988).

Fortunately, the situation for computer science does not tell the whole story. In the latter part of the 20th century, the expansion of computing and the Internet fueled the rise of applied IT fields in which technical skills, rather than being developed for their own sake, are increasingly put to use in the service of human needs. Applied fields, such as information science, information systems and instructional technology, have gained strength, and a new interdisciplinary field, informatics, has emerged. At the same time, interest in computer science itself is declining, especially among women (ITAA, 2005; Vegso, 2005). In this article, we explore the possibility that applied IT fields may provide more women-friendly cultures while still focused on technology. The larger question underlying this exploration is: Does applied IT education have the potential to bridge the “gender computing gap”?

BACKGROUND

Previous research has focused primarily on problems faced by women in computer science, especially in undergraduate-level education (Camp, 1997; Cohoon, 2001; Spertus, 1991). Many influences have been cited as contributing to the educational gender gap in computer science, including aspects of institutional culture, such as lack of role models (Pearl, Pollack, Riskin, Thomas, Wolf, & Wu, 1990), mentoring and advising (Whitely, Dougherty, & Dreher, 1991), informal networking (Smith-Lovin & McPherson, 1993) and a sense of belonging and identity (Ely, 1995); work-family conflict (Netemeyer, Boles, & McMurrian, 1996); and teaching styles (Cohoon, 2001; Turkle, 1988).

One of the most often-cited factors that discourage girls and women from studying computer science is the culture of computing itself: Computing is historically and conventionally associated with masculinity, an association that attracts boys to computers who then become role models for other boys, creating environments in which girls feel marginalized (Hacker, 1990; Turkle, 1988). Citing Margolis and Fisher (2002), Read (2002) argues that “women are further alienated by a stifling ‘geek culture’ that celebrates obsessive computing at the expense of broad interests.” Hackers are perceived as keeping exceptionally long hours and late nights, and manifesting “highly focused, almost obsessive behavior” (Frenkel, 1990, p. 38). For many women, this work ethic conflicts with their desire to start a family or, if they are older, their actual family responsibilities. Meanwhile, hackers, who are predominantly male, are seen as being “bright and creative,” while women remain on the periphery. Rasmussen and Hapnes (1991) suggest that this type of culture is important in producing and reproducing male domination in higher education in computer-related fields and that it influences the integration of women and their position within the field of computing.

The abstractness of much computer science instruction exacerbates the gender-biased culture, in that girls are more likely to be interested in real-world problem solving in contexts involving human users than in machines and programming languages (Clarke, 1992). A related deterrent for women is that most introductory computer science courses focus on programming skills rather than concepts of computer science. Girls and
women like IT, the findings suggest, but want to do something with it to improve the world; they are not satisfied with mastering computing skills for their own sake.

Applied fields such as information science, information systems and instructional technology have an a priori advantage over computer science when it comes to attracting future female professionals. They are grounded in the disciplines of real-world problems; for example, business, education and information management. Indeed, while women are significantly underrepresented in the IT workforce overall (ITAA, 2005), especially in computer science (Bryant & Vardi, 2002), applied careers, such as education and library and information science, attract higher proportions of women (e.g., Maata, 2003). However, research on the status of women in these professions shows mixed outcomes. On the one hand, women in applied disciplines are more likely to achieve professional parity or near parity with men (Quint, 1999; Wolverton, 1999). On the other hand, even in female-predominant applied professions such as education and library science, most high-ranking administrators are men (Growe & Montgomery, 2000; McDermott, 1998). Moreover, men tend to be disproportionately represented in IT work in these fields, which also tends to be more highly paid and more prestigious (e.g., Harris, 2000 for library and information science).

Little has been written about what leads women (or men) to choose applied IT careers, or what kinds of disciplinary cultures they encounter when they decide to study an applied IT domain. That is, there is no body of literature directly comparable to that available for computer science on which to base predictions about the long-term effects of the growth of applied IT fields on the gender computing gap. To address this research gap, we initiated the Information Technology Workforce (ITWF) project.

**INFORMATION TECHNOLOGY WORKFORCE PROJECT**

The ITWF project, funded by the National Science Foundation, has been collecting and analyzing data on the experiences, attitudes and outcomes of women and men in IT programs at five large public universities in the United States (U.S.) since September 2003. The focus is on students at all levels majoring in applied IT disciplines, specifically: information systems (IS), instructional systems technology (IST), informatics (I) and library and information science/studies (LIS).

We expected to find differences among these fields and between applied fields and computer science (CS); more generally, we expected that more female-oriented (“women-friendly”) disciplinary cultures would produce more successful outcomes (Jones, 1990), where success is defined both in terms of the quantity of women who enter IT-related educational programs and the quality of their educational experiences (Ahuja, Herring, Ogan, & Robinson, 2004).

Data collection for the project as a whole is via Web-based surveys, telephone interviews and face-to-face interviews with students, faculty and staff in the IT programs. In this article, we focus on findings from an initial Web survey of all IT student majors in the five universities (from the four applied IT fields plus computer science) that took place in spring 2004. The survey asked 100 questions about students’ experience, behaviors and attitudes regarding computers; their parents’ occupations and attitudes toward gender roles; student demographic information; and information about mentoring, stress and burnout in their academic environment.

**SELECTED FINDINGS**

The results of the survey were revealing of the attitudes and experiences that influence students in their selection of an IT career. Three distinct
patterns emerged, showing variation according to gender, IT discipline and the interaction between gender and discipline.

**Gender-Based Variation**

For some survey questions, men and women tended to answer differently, regardless of their program of IT study. For example, one question asked students to assess the relative importance they place on their career and their personal life. While most people answered that the two are equally important, females were more likely to choose the balanced response, while males were more likely to say that their careers are more important. This is consistent with previous observations that men in IT are more focused on their work, while women seek a more balanced lifestyle (Frenkel, 1990; Margolis & Fisher, 2002).

Another series of questions asked about work/life balance as a potential cause of stress. Most students answered “occasionally” or “seldom” to questions about whether schoolwork interferes with personal life and vice versa. However, women were more likely to say that school strain affects their personal life all the time, and men were more likely to respond that it never does. This could be a reflection of female students having greater domestic responsibilities (Netemeyer et al., 1996); more of the female than male survey respondents said they were married and/or had children at home.

Unsurprisingly, most of the IT students reported that they feel very comfortable using computers. However, here, too, a gender pattern is evident. Males in all majors were more likely to answer that they are very comfortable, whereas females gave most of the “somewhat comfortable” or “not at all comfortable” responses; this includes women in CS. Women also reported learning to use computers later than men, and playing fewer computer games when they were children.

These results support previous findings on gender differences in early computer experience and in computer self-efficacy (Durndell & Haag, 2002; Fromme, 2003), suggesting that some manifestations of the gender computing gap persist in applied IT disciplines.

**Discipline-Based Variation**

Differences are also evident within the IT disciplines, independent of respondent gender. A common tendency was for responses for LIS and IST to pattern together, in contrast to responses for CS and IS, with informatics falling in between. For example, when asked to weigh the relative importance of their career vs. their personal life, students in IS and CS were more likely to indicate that career was more important, and students in LIS and IST were more likely to say that their personal lives mattered more. Perhaps not coincidentally, students in IST and LIS also expressed greater satisfaction with their majors than did the IS and CS students.

Conversely, when asked what attracted them to their IT career choice, students in IS singled out earning a good living, and students in CS cited the challenge inherent in the subject matter. These considerations were rated less important by students in IST and LIS, for whom helping others was the highest consideration.

Disciplinary differences were also found with regard to students’ backgrounds and childhood experiences. CS and IS students were more likely to have had a father and/or mother who worked in an IT-related profession. CS, IS and informatics students clustered together in terms of where and when they first started using computers (at home, before or during their elementary school years) and the amount they played computer games as children (“very frequently” or “frequently”). Of those who answered that they did not learn to use a computer until high school or college, and played games “rarely” or “never,” most were from IST and LIS.

When asked what kind of job they expected to find after graduation, a majority of the re-
respondents, with the exception of IST students, answered “scientific or technical.” In addition, CS students preferentially indicated “university teaching and research,” and IS students indicated “administrative/managerial,” “consulting” and “sales.” IST students responded “consulting” and “K-12 teaching,” and LIS students indicated “administrative/managerial,” “clerical” and “other.” It is hardly surprising that different disciplines target different careers; however, the students’ career expectations in CS and IS (e.g., industry researcher, professor) tend to be higher status and better paid than those in IST and LIS (e.g., K-12 teaching, clerical). Computer science and business are traditionally “masculine” fields, whereas education and library science are traditionally considered “feminine” professions.

**Gender and Discipline Variation**

In the third pattern of survey results, gender and IT discipline interact. Specifically, males in traditionally masculine disciplines tend to pattern together with females in traditionally feminine disciplines, and females in traditionally masculine disciplines pattern together with males in traditionally feminine disciplines. These results suggest a natural mapping between the gender of the student and the traditional gender associations of the discipline.

For example, the survey asked students to indicate their degree of satisfaction with their major. While most answered “very satisfied,” CS men and IST/LIS women were overrepresented on that response. The inverse pattern was found for the less-enthusiastic response “somewhat satisfied”: CS women and IST/LIS men were overrepresented. In particular, women in IST and LIS selected “very satisfied” much more often than did women in CS, consistent with our expectations that IST and LIS are more “women-friendly” cultures (Jones, 1990).

Or consider another set of results, in response to the question, “How similar are your values to those of your field?” IS males and IST/LIS females were most likely to answer “very similar,” and informatics females and IST/LIS males were most likely to answer “not at all similar.” If we consider informatics to be part of the CS/IS cluster, this instantiates the same pattern.

Students’ confidence about finding a job after graduation displays a similar cross-distribution. Males in CS and females in IST/LIS were most likely to answer that it will be “very easy” to find a job; LIS males and male students in CS and informatics were most likely to answer “very hard,” and women in CS and informatics were overrepresented in the “don’t know” category. There seem to be two types of male CS student; those who are very confident and those who are more pessimistic about their job prospects. Apart from this, these results also fit the tendency for confidence and satisfaction to accompany a mapping of student gender onto IT discipline.

**DISCUSSION**

The results of a Web-based survey of students in different IT programs in five U.S. universities reveal gender differences among students, differences among IT disciplines and a mapping of gender onto disciplinary cultures, all of which potentially affect students’ choice of and satisfaction with an IT career. Rather than clustering together in contrast to CS, the responses of students in applied IT programs tended to fall out along gender and disciplinary lines, with CS and IS clustering at one end, and LIS and IST at the other; informatics, the new discipline, mostly fell in between.

Somewhat discouragingly, the women in the applied IT programs, like the women in computer science, report having less early computer experience and less confidence in their computing abilities than their male peers (Ogan, Ahuja, Robinson, & Herring, in press). At the same time, the female students’ reported preference for a
balanced lifestyle may help to explain why more women choose applied IT majors than CS. The most popular IT majors among women in our study are those that are traditionally feminine, in which there are more women peers and role models, and which value using technology to help people. Fewer women choose IT majors where the focus is on technology for its own sake (in the case of CS) or making money (in the case of IS), although the women who choose those majors appear to do so for those reasons, similar to men. Most intriguingly, both genders report greater satisfaction with and optimism about their career choices in IT disciplines that mirror the gendered cultures found in society at large.

**FUTURE TRENDS**

These findings are simultaneously encouraging and problematic. On the one hand, they suggest that the rise of applied IT fields does indeed open the door to more women studying and working in IT. Specifically, the findings indicate that IT programs located in traditionally “women-friendly” contexts may be more accessible and appealing to women, and by extension, that other IT contexts might evolve to become more women-friendly as well. Informatics is an example of a discipline whose culture has not yet become fixed; in this study, it tended to pattern weakly with CS and IS, but that tendency could shift.

At the same time, the survey findings are problematic in that they re-inscribe cultural stereotypes about gender roles and interests in the IT domain. Moreover, a positive interpretation of the results assumes an equivalence between the various IT fields that is more illusory than real. “Feminized” professions such as instructional design and librarianship still have less status and are lower paid, even when they incorporate IT (Lorenzen, 2002). Moreover, the computing technology involved is typically less rigorous than in computer science (e.g., it may not require practitioners to know how to program a computer). Does women’s greater satisfaction with applied IT programs help to bridge the gender IT gap, or does it reproduce a larger societal status quo, of women as less technology- and career-oriented?

Ideally, IT study should be equally accessible and rewarding for women interested in hard-core computing and men interested in human applications as it is for men interested in, for example, grid structures, or women interested in information management or educational applications. Again, informatics may help to bridge the gap if a culture can be forged in which technical rigor is balanced with a focus on computing applications that benefit people.

**CONCLUSION**

This study has raised a previously unasked question about the potential of applied IT fields to close the gender computing gap. The question remains, for the most part, unanswered; further research needs to be done, for example, to compare the IT content of applied IT programs, and to take national culture (in addition to disciplinary culture) into account in assessing the appeal and outcomes of different approaches to IT. In the meantime, enrollments in CS are declining in U.S. universities (Vegso, 2005) and the IT workforce is aging (ITAA, 2005), but society’s need for IT expertise remains high. In the future, it is likely that career seekers of both genders will turn increasingly to applied IT and to the new interdisciplinary paradigms emerging from combinations of computing, information studies and other fields (Berghel & Sallach, 2004). It is important to understand the social dynamics and implications of these trends, in the hope that we may be able to shape them into gender-equitable configurations.
REFERENCES


**KEY TERMS**

**Applied IT:** Programs of study and careers in which information technology is applied in contexts of use; for example, business, education, libraries.

**Computer Science (CS):** In this article, a university program teaching computer science.
Gender Computing Gap: The relative absence of women in computing-related programs of study and careers, especially computer science.

Informatics (I): A new interdisciplinary area of study in the U.S., with roots in computing applications to medicine, and university programs that teach it.

Information Systems (IS): In this article, a university program teaching information systems; usually associated with Management in Business Schools.

Instructional Technology (IT): In this article, a university program teaching instructional technology; typically found in Schools of Education.

Library and Information Science/Studies (LIS): In this article, an academic program teaching library science, information science and/or other information management skills.

ENDNOTES

1 National Science Foundation IT Workforce Grant #0305859, “Toward Gender Equitable Outcomes in Higher Education: Beyond Computer Science.”
2 Universities were selected that offered computer science plus a minimum of two applied IT programs.
3 1,768 students responded to the survey, of whom 1,516 indicated their gender.
4 A “finding” in this article indicates a result for which a gender or discipline had values well above the mean for the specified discipline, based on descriptive statistics. Statistical significance was not measured and no claims of significant differences are made.
5 See Ogan et al. (in press) for more detailed discussion of gender differences in the survey responses.
6 Women are seriously underrepresented in both business information systems and computer science programs in our sample.

This work was previously published in Encyclopedia of Gender and Information Technology, edited by E. Trauth, pp. 474-481, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
INTRODUCTION

While traditional face-to-face (FtF) forms of interaction have proven disadvantageous to females in mixed-sex settings, computer-mediated communication (CMC) holds the promise of helping to level the playing field between the sexes, at least in terms of equitable communication between genders. However, evidence from recent research shows that gender inequalities persist. The objective of this article is to shed light on why the promise of gender equalization in CMC is not evidenced.

BACKGROUND

The equalization phenomenon of CMC is attributed to the reduction in social cues associated with an online virtual environment (Dubrovsky, Kiesler, & Sethna, 1991; Kiesler, Siegel, & McGuire, 1984; Kiesler, Zubrow, Moses, & Geller, 1985; Siegel, Dubrovsky, Kiesler, & McGuire, 1986). In this comparatively lean mode of interaction, it is argued, status cues are filtered out, which leaves people feeling more anonymous and less individual. This “deindividuation” lowers self-awareness and self-regulation, resulting in less evaluation apprehension and overall reduced social inhibitions. According to Kiesler et al., the relative distance and anonymity afforded by CMC decreases the salience of status, resulting in the increased participation of lower status members. The emphasis shifts from message contributor to message content, thus serving to equalize the influence of high-status individuals.

Indeed, a repeated finding in the early research in this area is that both participation in group discussion as well as influence over group outcome is more equal under conditions of asynchronous CMC compared to traditional FtF interaction (Clapper, McLean, & Watson, 1991; George, Easton, Nunamaker, & Northcraft, 1990; Hiltz, Johnson, & Turoff, 1986; McLeod, 1992;
Gender Equalization in Computer-Mediated Communication

Rice, 1984; Zigurs, Poole, & DeSanctis, 1988). Females, being of lower status than males (in most cultures), are thus expected to fair better in a CMC context as compared to a traditional FtF context. However, as pointed out by Postmes and Spears (2002), many studies of electronic media use have not found evidence to support the equalization hypothesis (Adrianson & Hjelmquist, 1991; Berdahl & Craig, 1996; Hollingshead, 1996; Matheson, 1991; Saunders, Robey, & Vaverek, 1994; Straus, 1996; Weisband, 1994; Weisband, Schneider, & Connolly, 1995).

REVIEW OF LITERATURE

The article begins by reviewing literature on status, particularly as it relates to gender differences in traditional FtF communication environments before moving into CMC environments. The focus of the review concerns mixed-sex, task-oriented work situations.

Sociological-Based Theories

Sociological-based theories pertaining to status include status-characteristics theory (SCT) and social-role theory. SCT is concerned with the effects on face-to-face interaction of differences in individuals’ status. A central tenet of SCT is that status hierarchies influence interaction in groups (see Wagner & Berger, 1993, 1997, for summaries). Findings indicate that high-status members contribute more opinions and enjoy increased influence in groups. Regarding gender, males are accorded a higher status than females and are believed to have more expertise overall (e.g., Eagly & Carli, 1981; Kent & Moss, 1994; Wood & Karten, 1986). Thus, both sexes expect higher task performance from males, independent of whether gender is relevant to the group’s task at hand (Berger, Rosenholtz, & Zelditch, 1980). In mixed-sex groups, females contribute less task-relevant content due to the expectation of superior male performance. As pointed out by Johnson et al. (1998), these findings are consistently supported (Anderson & Blanchard, 1982; Dovidio, Brown, Heltman, Ellyson, & Keating, 1988; Lockheed & Hall, 1976; Piliavin & Martin, 1978; Strodbeck & Mann, 1956; Woody & Karten, 1986).

The SCT research considers traditional FtF interaction. In an attempt to tease out the effect of differing degrees of FtF interaction on the effects of status, Mueller et al. (2002) found that the predictions from SCT are more likely to be supported when “women and men regularly interact face-to-face” (p. 178). Thus, they conclude that the amount of face-to-face interaction best predicts whether SCT’s claims on gender inequalities will be supported.

Social-role theory (Eagly, 1987) asserts that males and females are socialized differently such that each sex learns dissimilar (i.e., gender-appropriate) behavioral patterns. Females are socialized to respect and defer to males and to exhibit relative docile behavior (Seibert & Grunfeld, 1992). Males, on the other hand, are socialized to be more assertive, competitive, and aggressive (Eagly & Steffen, 1984; Powell, 1988).

These socialization processes result in individuals exhibiting stereotypical traits and behavior associated with their gender, which is reflected within the interaction of participants in mixed-sex groups (Broverman et al., 1972; Eagly & Steffen, 1984; Strodbeck & Mann, 1956). This body of research generally finds that males participate more and are more influential in mixed-group settings than their female counterparts (e.g., Eagly & Carli, 1981; Williams, 1992). Males also emerge more frequently as group leaders (Eagly, 1987). These findings hold regardless of the sexual composition of the group (see below). Although social roles have become less rigid over the years, gender stereotypes continue to persist (Biernart & Wortman, 1991; Diekma & Eagly, 2000; Steil, 1997).
Structural Theory

The theory of proportional representation (Kanter 1977a, 1977b) provides a structural approach for accounting for within-group behavior due to status differences. It posits that the numerical representation of a status category (e.g., race, sex) influences intragroup interaction. A group member from the numerical minority experiences feelings of isolation and powerlessness. This leads to behavior by the numerical minority that tends toward passive and inhibited conduct. As a means of lessening the feelings of isolation and powerlessness and to fit in, the numerical minority may adopt the behavioral characteristics of the numerical majority. These behaviors are evident in “tilted” groups, where group members account for between 15% to 35% of the minority status, but are more prevalent in “skewed” groups, where they represent less than 15% of group membership. Polarization occurs as the numerical majority alienates the numerical minority by discounting contributions of the minority.

In terms of mixed-gender groups, proportional-representation theory suggests that the numerical representation of men and women directly influences behavior rather than the sex or socialized-gender roles of the individuals themselves. Research suggests that the results of proportional representation are quite direct. For example, Johnson and Schulman (1989) found that solo female members engaged in task activities significantly below the group average. However, there is evidence that men and women are differentially affected by underrepresentation. That is, when females are in the majority and there is a lone male group member, the solo male may dominate (e.g., Crocker & McGraw, 1984; Williams, 1992).

Linguistic Differences in Communication of Men and Women

Researchers within the area of sociolinguistics have uncovered numerous differences in the way males and females use language to communicate and interact. These differences occur to such an extent that sex-specific patterns of communication are evident in discourse (Coates, 1986; Preisler, 1987). Men’s discourse is more competitive and involves preserving their independence, while women’s is supportive, consensus seeking, and socially oriented (Coates; Preisler; Rhodes & Wood, 1990). Women express more agreement and seek the opinions of others to a greater extent than men (Eakins & Eakins, 1978).

In both mixed and homogeneous groups, men contribute more task communication (e.g., Bartol & Butterfield, 1976; Ellis & McCallister, 1980; Hare, 1976; Nemeth, Endicott, & Wachtler, 1976; Strodbeck & Mann, 1956; Sturm, 1989). Women, however, contribute more to statements relating to group maintenance (e.g., Bartol & Butterfield; Ellis & McCallister, 1980; Strodbeck & Mann; Sturm). While both task and group-maintenance communication is necessary for effective group functioning (Bales, 1950), leadership is consistently attributed to those who contribute task, rather than maintenance, input (e.g., Hare; Nemeth et al.; Strodbeck, James, & Hawkins, 1957).

Specific linguistic features are also associated with each sex. Females use more expressive and emotional language, and their language incorporates a variety of markers that males do not tend to use (Lakoff, 1975). These markers include hedges (e.g., “I’m not sure, but...”), (very) polite forms (e.g., “I’d really appreciate it if...”), tag questions (e.g., “You’re leaving now, aren’t you?”), direct quotes (men paraphrase more often), and apologies (e.g., “I’m sorry, but I feel that...”). Women also speak less frequently than men in mixed-sex settings.
Gender-EQUALIZATION Studies in Asynchronous CMC

While there is a body of research pertaining to equalization and status in CMC (see above), the number of studies specifically relating to gender equality in CMC is quite sparse. Studies purporting the status-equalizing effects of CMC do so based on non-gender-related status markers (e.g., occupation, Saunders et al., 1994; education, Weisband et al., 1995). By and large, CMC research has not studied gender equalization or controlled for gender effects. Below, I review CMC studies relating to gender equalization.

Selfe and Meyer (1991), in one of the earliest studies to report on gender in CMC, found that males and other high-status members dominated online interaction under normal as well as anonymous conditions. Herring has conducted a series of studies on gender in CMC. In a particularly telling study, Herring (1992, 1993) conducted a 5-month-long analysis of two academic Listservs. In her study of their discourse, she identified linguistic features associated with each gender. These features comprised significant sex-based differences, such that they formed a style recognizable as woman’s language and men’s language. The features of women’s language included explicit justifications, questions, apologies, personal orientation, and support of others. In contrast, men’s language included strong assertions, self-promotion, rhetorical questions, authoritative orientation, humor, and sarcasm. These findings are similar to the aforementioned linguistic differences noted in traditional FtF interaction.

Furthermore, in terms of participation, Herring (1993) found that a minority of men dominated the discourse for both the amount and content of contributions. She notes that “the most striking sex-based disparity is the extent to which men participate more than women” (p. 4). Although women constituted between 36% and 42% of the membership on the Listservs, their communication rate was significantly less than what would correspond to their numerical presence. Herring goes on to relay that when women attempted to participate on a more equal basis, they were actively censored by men who either ignored their contributions or attempted to delegitimize their postings. Herring observes that because women are conditioned to avoid conflict, they avoided participating. She concludes that

[r]ather than being democratic, academic CMC is power-based and hierarchical. This state of affairs cannot however be attributed to the influence of computer communication technology; rather, it continues pre-existing patterns of hierarchy and male dominance in academia more generally, and in society as a whole. (p. 11)

Recently, Postmes and Spears (2002) conducted a series of laboratory studies on equalization in groups communicating via CMC by looking at dominance, the antithesis of equalization. All teams consisted of four undergraduate students and were balanced in terms of sex (i.e., two females, two males). They analyzed dominance in terms of the content of contribution, and also counted the number of words and postings of the participants. While their studies did not show evidence of male dominance in terms of word or statement counts, males were found to dominate team discussions in terms of linguistic style. Males used forceful, independent, confident, and assertive language in order to dominate. Furthermore, under conditions of supposed anonymity, gender differences were more salient. They attribute this to the inability to be truly anonymous when communicating. Due to patterned differences in the discourse of men and women, it is thus likely that gender can be determined even under anonymous conditions where gender is hidden.

Postmes and Spears (2002) conclude that the prior mixed findings regarding equalization can be attributed, at least in part, to employing an inappropriate measure based on a numerical count. Their focus on measuring content differs from
previously cited studies that found equalization in CMC, which relied on measures of equalization based on the numeric count of words and postings. Postmes and Spears warn that researchers must consider the content of the contributions to determine a true measure of the equality of participation and influence.

FUTURE TRENDS

There is certainly room for more gender studies in CMC. While studies by linguists (Herring in particular) have informed our understanding of gender differences in CMC in terms of discourse, these field studies have largely been conducted using discussion boards. While other studies of status and equalization have been conducted using task-oriented teams, most have not looked at gender. And while there are some studies of gender on teams interacting via CMC, these do not consider status and equalization. Finally, there are no studies that focus on the gender composition of teams as an independent variable.

CONCLUSION

What can be concluded from the limited research is that CMC does not promote gender equalization. Many of the conditions leading to gender inequality in traditional FtF interaction find their way into the context of CMC. As described by sociological theories as well as structural theories, status effects relating to gender appear to carry over into CMC. However, significant gaps in gender research in CMC remain.

REFERENCES


Lockheed, E., & Hall, K. (1976). Conceptualizing sex as a status characteristic: Applications to


Woody, W., & Karten, S. (1986). Sex differences in interaction style as a product of perceived sex


**KEY TERMS**

**Computer-Mediated Communication (CMC):** Electronic communication using a computer.

**Deindividuation:** The loss of individual identity and a gaining of the social identity of a group.

**Equalization Phenomenon:** The salience of status is decreased by interacting in a CMC environment; it is a hypothetical result.

**Expectation-States Theory:** Argues that power and prestige differences within a group arise from differing performance expectations that members hold for themselves and others.

**Heterogeneous Group:** Mixed-sex group (i.e., male and female group members).

**Homogeneous Group:** Same-sex group (i.e., all female or all male).

**Lean Mode of Interaction:** Due to the reduction in social cues, CMC is considered lean compared to face-to-face interaction.

**Social-Role Theory:** Asserts that males and females are socialized differently such that each sex learns dissimilar (i.e., gender-appropriate) behavioral patterns.

**Status-Characteristics Theory (SCT):** Is concerned with the effects on face-to-face interaction of differences in individuals’ status; a central tenet of SCT is that status hierarchies influence interaction in groups.

**Theory of Proportional Representation:** Provides a structural approach for accounting for within-group behavior due to status differences; it posits that the numerical representation of a status category (e.g., race, sex) influences intragroup interaction.
Chapter 5.4
The Cross-Cultural Dimension of Gender and Information Technology

Haiyan Huang
The Pennsylvania State University, USA

INTRODUCTION

The underrepresentation of women in the information technology (IT) sector has been widely studied in the contexts of western countries such as North American and Western European countries. These studies indicate that the underrepresentation of women in the IT sector is unveiled in multiple perspectives, including IT access, the development, adoption and use of IT, IT education, and the IT workforce in general and within the IT workforce structure itself (e.g., different levels of IT positions) (Cooper & Weaver, 2003; Gürer & Camp, 2002; Hartzel, 2003; Klein, Jiang & Tesch, 2002; Margolis & Fisher, 2002; Rommes, 2002; Trauth, 2002; von Hellens, Neilsen, & Beekhuizen, 2001; Webster, 1996).

Why is it important to study issues related to gender and information technology? First, it is argued that the information technology sector should value and leverage all kinds of diversity (including gender diversity as one dimension) to enhance productivity, to facilitate IT innovation, and to develop IT for a wide variety of people (Avgerou, 2002; Roberts, 2003; Trauth, Huang, Morgan, Quesenberry, & Yeo, 2006). Second, it is also argued that women’s underrepresentation in and exclusion from information technology can be attributed to power and socio-cultural reproduction of inequality through technology development and use, and the historically social-construction of technology fields as “masculine” domains, which result in a gendered digital divide (Cockburn, 1985; Kvasny & Trauth, 2002; Kvasny & Truex, 2001; Wajcman, 1991, 2004; Woodfield, 2000). Ignorance or failure to address issues related to gender and IT will further marginalize women’s participation in future economic and social development, and will endanger social equality and social welfare in general (Kvasny & Trauth, 2002).

A significant trend of the contemporary information technology industry is towards globalization, which is manifested through a
variety of established practices such as IT offshore outsourcing, global software development, and innovation through global R&D (research & design) collaboration (Sahay, Nicholson, & Krishna, 2003; Walsham, 2000, 2001, 2002). Such a globalization trend of the IT industry and market has put forward new challenges to gender and IT research, to incorporate the cross-cultural dimension. Similar to the rationale for studying gender and IT in developed countries (leveraging diversity and improving social inclusion), Hafkin and Taggart (2001) argued that it is imperative to examine the cultural factors while studying gender and IT in developing countries.

Although the research on the cross-cultural dimension of gender relations with information technology is limited, Galpin (2002) pointed out that the underrepresentation of women in IT seems to be a worldwide phenomenon indicated by statistics. Galpin (2002) also pointed out that there is a wide range of participation in IT by women, which is influenced by complex cultural and societal factors that are different from country to country. Models related to gender relations to IT developed in certain socio-cultural contexts may not be applicable to others (Clarke & Teague, 1994a; El Louadi & Everard, 2005; Mukhopadhyay, 1996, 2004).

In addition to the importance of studying gender and IT within a specific cultural context, there is another perspective of the cross-cultural dimension of gender and IT: the increasingly diversified global IT workforce as a result of the IT skill shortage, global IT outsourcing, and other global IT collaborations. For example, under the pressure of the skill shortage of the IT workforce in America, the IT institutions and industry turn to the global intellectual pool for recruiting talented international students and skilled IT workers through F-1 and H1-B visas (National Research Council, 2001). According to the 2005 ITAA (Information Technology Association of America) report, the representation of Asian IT workers in the IT workforce doubled the number of Asian worker in the overall workforce in America (ITAA, 2005). Globally, an increasing number of countries have a maturing IT sector, which enables them to enter the global IT outsourcing market or to engage in globally distributed collaborative software work through virtual environment (Trauth et al., 2005). The mobility and cultural diversity of the global IT workforce will add more complexity to articulate their gender relations to IT since individuals may have different cultural backgrounds and experiences.

Therefore, it is important for scholars and practitioners to explore the cross-cultural dimension of gender and IT to understand how the nuances of different cultural influences shape women’s relations with IT, to build knowledge with respect to the plural perspective of gender and IT research, and to prepare future global IT workforce. In this article, I first articulate the theoretical underpinnings of cross-cultural dimension of gender and IT research, then review some current studies related to this research area, and finally discuss some future research agendas.

BACKGROUND

The theoretical explanations about why women are underrepresented in IT education and the workforce are complex in nature (Adam, Howcroft, & Richardson, 2002, 2004; Webster, 1996). Those theories attempt to articulate the gender IT gap from either the technology perspective (including the IT sector environment) or the people perspective (e.g., women). From the technology perspective, there are two theoretical orientations: technology determinism and social shaping of technology. From the people perspective, there are three theoretical orientations: essentialist theory, social construction theory, and Individual Difference Theory of Gender and IT (Trauth, Quesenberry, & Morgan, 2004).

Technology determinism views technology as a value-free artifact without any biases, and the
The Cross-Cultural Dimension of Gender and Information Technology

impacts of technology on society are self-directed and universal. It neglects the influences of socio-cultural contexts in which the technology is created and used. Social shaping of technology, on the other hand, views technology as a product of socio-cultural and power relations and the impacts of technology on society are shaped by the socio-cultural beliefs embedded within the technology and the societal attitude towards the technology. For example, Kvasny and Truex (2000) argued that power relations and controls are embedded in technology and in society’s attitudes towards technology, and hence new technology tends to “reify the dominant relations in the existing social order.” The embeddedness of social order in technology will privilege certain groups while disempowering others (Kvasny & Trauth, 2002). Technology related educational disciplines and industrial sectors have been historically dominated by men and hence are embedded with belief systems that are engendered with “masculinity,” which may be exclusive of women’s participation (Cockburn, 1985; von Hellens et al., 2004; Wajcman, 1991, 2004). Driven by this theoretical orientation, researchers recommend bringing IT closer to women: changing technology development methods, increasing women’s participation in system development and implementation, modifying the “masculine” culture of IT at various levels from societies, to industries, and to organizations and educational institutions.

The essentialist theory argues that men and women are differentiated by their biological differences; such differences are inherent and fixed and thus act as main causes to determine their different relationships with IT (Adam et al., 2004; Trauth et al., 2004). Similar to the technology determinism, the essentialist theory only emphasizes the intrinsic internal differences between men and women and overlooks the influences of social contexts in which both technology and people are embedded. Social construction theory views gender identity and gender relationship with IT as on-going negotiation processes in social interactions, in which the influences of various factors are relevant and important (Trauth et al., 2004; von Hellens et al., 2004; Wajcman, 1991; Webster, 1996). Driven by this theoretical orientation, researchers recommend bringing women closer to IT, which include setting up mentoring programs, providing convenient child-care facilities, and developing initiatives to facilitate work-life balance of female IT workers.

Trauth (2002) pointed out that social construction theory of gender and IT tends to separate men and women as two separate socio-cultural groups, which mainly focuses on the differences between the gender groups while overlooks the differences within the gender group. Within the gender group, differences such as ethnicity background, family, social class, individual life experiences, age, individual personality and attitude also play important roles in shaping one’s relationship with IT (Kvasny, 2003; Trauth et al., 2004). Trauth argued for focusing on individual level of analysis and developing Individual Difference Theory of Gender and IT (Trauth, 2002; Trauth et al., 2004).

MAIN THRUST OF THE ARTICLE

Theoretical Underpinnings of Cross-Cultural Dimension of Gender and IT

In his research on cross-cultural information systems production, use, and management, Walsham (2001, 2002) emphasized the importance of examining the cultural diversity and argued that the existing local socio-cultural context is a critical factor in mediating the globalization process in a specific context and, in turn, will have an impact on the complexity of globalization. Similarly, to study the cross-cultural dimension of gender and IT, the underlying theories should be sensitive to and account for the dynamics of cultural diversity. Evidences show that cultural diversity and its influences on gender and IT is connected to
three theoretical perspectives: social construction of gender relationship with IT, social shaping of technology, and Individual Difference Theory of Gender and IT.

El Louadi and Everard (2005) examined current labor force condition of Arab countries, and more particularly Arab women. They pointed out that because of the Arab cultural influences, the future opportunity of engaging Arab women in IT work may lie in telework, which can bring work and information technology into the home where Arab women can be professionals and care for the family at the same time. This is an example of the distinct Arab socio-cultural influence on women’s relationship with IT.

Nielsen et al. (1998) studied non-Asian and Asian female students enrolled in IT disciplines in Australia and the differences among their perceptions of IT educations and careers. Their study showed that both non-Asian and Asian females had similar views about the IT professions. However, Asian female students were more inclined to choose IT related subjects despite their negative perceptions of IT professions because of the future prospects of employment opportunities (Nielsen, von Hellens, Greenhill & Pringle, 1998; Nielsen, von Hellens, Pringle & Greenhill, 1999). Such a choice is influenced by collectivist characteristic of Asian culture and is based on practical consideration, as compared to the “free-choice” decision making in most Western cultures. The research by Geary on comparing children’s mathematic achievements in different countries also supports this argument (Geary, 1996; Geary, Bow-Thomas, Liu & Siegler, 1996): Asian culture emphasizes practical reason (e.g., usefulness) in decision-making while Western culture emphasizes free choice. The studies by Nielsen et al. (1998, 1999) indicate that although the perceptions of IT professions might be similar between Asian and non-Asian women (e.g., the social shaping of technology is similar), their relationships with IT are different because of the cultural influences. Also their studies are situated in one national context, Australia, which implies that researchers should pay close attention to the differences within gender groups (in this case, Asian and non-Asian Australian female IT students). It also implies that cross-cultural influence is an important dimension in studying the Individual Difference Theory of Gender and IT.

An early study by Clarke and Teague (1994a, b) is also about the Australian context. Based on interview data, their study shows that Asian female CSE (Computer Sciences and Engineering) students did not see computing as a male domain and they also received direct encouragement from family to pursue computing related secondary studies. This is an example of cultural influences on both social shaping of IT and social construction of gender relationship with IT. Clarke and Teague (1994b) pointed out that “the differences within gender groups are greater than the differences between gender groups” (p. 259), which indicates the relevance and important in studying the individual difference of gender and IT.

The study by Mukhopadhyay (1996, 2004) also strongly supports considering a cross-cultural dimension in studying gender and science (although IT is not a central subject in this study, it is closely relevant). Mukhopadhyay (1996, 2004) challenged both the essentialist theory and those American theories of gendered science that embodied long-standing American cultural models to other socio-cultural contexts. She pointed out that it is the Indian cultural model of family, gender, and science that influences the Indian women’s academic decision-making process, and hence results in a predominantly male scientific community in India. Mukhopadhyay (2004) also stressed that:

“We need to explore the processes through which individuals internalize and utilize what Strauss and Quinn (1997) call “public culture” as it relates to gender, schooling, academic subjects (mathematics, science, etc.), family, kinship, and marriage. We need to understand more about how...
and why some messages have more motivational force than others, for some girls more than others, and how they mediate math and science attitudes, achievements, and choices. (p. 487)

**Current Research on Cross-Cultural Dimension of Gender and IT**

A comprehensive review of current research on cross-cultural dimension of gender and IT is beyond the scope of this article. Here, I would like to provide a brief overview. Current research on cross-cultural dimension of gender and IT can be classified into two categories. One category consists of comparison studies that investigate how and why the women's participations in IT are different in different countries. The study by Galpin (2002) provided a good summary and references for this type of study. Another example is the study by Hafkin and Taggart (2001), which focuses on studying the developing countries. The primary objectives of those studies are: to understand the nuances of different socio-cultural influences on gender and IT in different countries; to come up with solutions to improve women's participation in IT that fit with a specific socio-cultural context; and to see whether or not some practices and experiences can be transferred from one country to another country.

Another type is to situate the cross-cultural dimension of gender and IT within a specific societal context while studying how the diverse cultural backgrounds of different women may influence their relationships with IT. The studies by Clarke and Teague (1994a), Nielsen et al. (1998, 1999), and by Mukhopadhyay (1996, 2004) are the examples. The primary objectives of these studies are to critically address the cultural influences on women's relationships with IT within gender groups, and to leverage the influences of cultural diversity on IT to a great extent. The number of this type of study is limited compared to previous types. However, the importance of this type of research is increasing due to the globalization of the IT market and industry and the mobility of IT professions.

**FUTURE TRENDS AND CONCLUSION**

In this article, I argue that it is important to investigate the cross-cultural dimension of gender and IT because of the contemporary globalization phenomena. From the theoretical perspective, I discuss how the cross-cultural dimension can be incorporated into the social shaping of technology, social construction of gender and IT, and individual differences perspective on gender and IT. However, the number of studies in cross-cultural dimension of gender and IT is limited.

Therefore, one of the future research agenda items is to have more studies in this area to address the influences of cultural diversity. Another future research agenda item is to further incorporate the cross-cultural dimension into the theoretical developments of gender and IT. To do so, we need to encourage interdisciplinary discourse among different disciplines, and also global collaborations in gender and IT research.

**REFERENCES**


**KEY TERMS**

**Essentialism:** The assertion of fixed, unified and opposed female and male natures (Wajcman, 1991, p. 9). This theoretical perspective is used to explain the underrepresentation of women in the information technology field by arguing the existence of “essential differences” between males and females with respect to engagement with information technology.

**Individual Differences Theory of Gender and IT:** A social theory developed by Trauth (Trauth, 2002; Trauth et al., 2004) that focuses on within-group rather than between-group differences to explain differences in male and female relationships with information technology and IT careers. This theory posits that the underrepresentation of women in IT can best be explained by considering individual characteristics and individual influences that result in individual and varied responses to generalized environmental influences on women.

**Social Construction:** A theoretical perspective articulated by Peter Berger and Thomas Luckmann (1967) that focuses on social processes and interactions in the shaping of actors. This theoretical perspective is used to explain the underrepresentation of women in the information technology field by arguing that technology—socially constructed as a masculine domain—is in conflict with socially constructed feminine identity.

*This work was previously published in Encyclopedia of Gender and Information Technology, edited by E. Trauth, pp. 147-153, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
INTRODUCTION

Prior to the Internet, forms of social expression, communication, and collaborative behavior are known to be sensitive to cultural nuances. According to researcher Geert Hofstede (1991), a widely used definition of culture is proposed where “Every person carries within him or herself patterns of thinking, feeling, and potential acting which were learned through their lifetime” (p. 4). Hofstede referred to such patterns as mental programs or “software of the mind.” It is expected that such mental programming related to cultural differences will affect perceptions of the electronic medium as well (Raman & Watson, 1994). Related to the topic of this volume, culture has a place in the consideration of e-collaboration when individuals come together to work toward a common goal using electronic technologies. This may include various domains including e-business, e-learning, distributed project management, working in virtual teams of various forms, to name a few.

While there is little work to date on the explicit topic of culture and e-collaboration, there is evidence that creating culturally appropriate user interfaces (Cyr & Trevor-Smith, 2004) contributes to a better perception of the interface (Kondratova & Goldfarb, 2005), and indeed to enhanced levels of Web site trust and satisfaction (Cyr, 2006). In e-commerce settings, Web sites that are perceived as appropriate to the user have also resulted in greater commitment (Oliver, 1999). In this article, and building on previous work in related areas, it is argued that the development of culturally appropriate electronic interfaces can enhance user involvement, ultimately resulting in enhanced e-collaboration.

In the following sections, culture as a context for e-collaboration is outlined followed by considerations of the Web used as a communication tool, and how trust and satisfaction are related to
the online collaborative process. The article ends with concluding remarks.

**CULTURE AS CONTEXT**

Over the decades sociologists have proposed that socially shared meanings are culture specific. These shared meanings are grounded in language, geographical proximity, and history as shared by members of nations or those who have lived within the same social environment (Hofstede, 1980). In both commercial and noncommercial settings, culture has been found to have implications for information systems research. More specifically, culture is proposed to affect online trust (Jarvenpaa, Knoll, & Leidner 1999), Web site development (Sun, 2001), use of group support systems (Reining & Mejias, 2003), predisposition for type of electronic communication media (Straub, 1994), among other topics.

Related to how individuals operate together in groups, mental schemas for knowledge construction (Kock, 2004) influence the impact of e-collaboration technologies on the individuals involved. Schemas can be socially constructed, causing groups to interpret information in specified ways (Lee, 1994)—thus influencing perception and interaction. As Kock (2004) elaborates, the degree to which members of a task group share similar schemas, then less cognitive effort is required to successfully accomplish the task. Members of the same cultural group are more likely to share similar mental maps or schemas than with members of external groups.

In decision making, Tseng and Stern (1996) found significant differences in the information gathering behavior between Asians and North Americans. Different online communication strategies were uncovered in a study that included Japan, Spain, and the United States (Okayazaki & Rivas, 2002). Further, cultural differences exist in instant messaging between Asia and North America (Kayan, Fussell, & Setlock, 2006). For example, North Americans reported significantly less multiparty chat and rated emoticons lower in importance than Asians. Ethnicity has been established as a factor in electronic brainstorming (Tan, Wei, Watson, Clapper, & McLean, 1998). Based on the foregoing, it is a natural extension that culture influences e-collaboration behavior.

**THE WEB AS A COMMUNICATION MEDIUM**

Bordia claims that “computers and electronic networks have revolutionized communication” (1997, p. 99). Although it is not always clear that the electronic medium has enhanced the communication process since nonverbal cues that form a significant portion of the transmitted message are mostly missing. In fact, research has demonstrated that an absence of nonverbal cues that serve to “embellish meaning or social context regarding gender, age or status” can potentially hamper communication efficiency (Bordia, 1997, p. 9). Alternately, new capabilities for communicating content and collaborating using the Web are created (Tsao & Lin, 2001).

Simon (2001) used media richness theory (from Daft & Lengel, 1986) to examine how information richness might enhance user perceptions of the interface. Various design characteristics of Web sites were considered such as shapes, colors, language, site layout, and quality of information. It was expected that information rich Web sites would reduce user ambiguity, increase trust, reduce perception of risk, and encourage users to utilize the site. Significant differences between cultural groups in the study were uncovered, with Asians registering higher levels of trust with information across all Web sites in the study than European and North American groups. This finding suggests that not only is creating information rich interfaces useful generally, but that across cultures different preferences for the user interface prevail.
In studies on graphical user interfaces (GUI) in group support systems (GSS), benefits of using icons over text have been established, although this research does not explore aesthetics or usability of the system (Sia, Tan, & Wei, 1997). One would expect that computer supported, interactive, visual representations contribute to a user’s assessment of the medium as more effective, and may vary across cultures. As already noted, in one study on instant messaging differences in perceptions of emoticons across cultures were discovered. In other work, German users were found to value hierarchy and verbal components of a Web page, while Asia users prefer visuals (Sun, 2001). Color connotes different cultural meaning (Barber & Badre, 2001; Cyr & Trevor-Smith, 2004). Red means happiness in China but danger in the United States. In a study that compared Canadian, American, German and Japanese users, Japanese favored a more visual and “emotional” approach to user interface design (Cyr et al., 2005).

Related to the preceding, and in the context of e-collaboration, it is expected that if user interfaces are culturally appropriate then users are more likely to become engaged in the communication process, and with one another. This would be the case whether cross-cultural collaboration occurs in e-business, e-learning, or distributed work settings.

**TRUST AND E-COLLABORATION**

The use of computer-mediated and Web-based communication technologies has created new virtual environments in which trust plays a significant role. Online trust between collaborating parties is more difficult to elicit than trust in traditional settings since users have fewer cues than in a face-to-face or brick-and-mortar context. The primary communication interface for the user is an information technology artifact rather than a person.

Despite the complexity involved in studying online trust, in recent years a body of knowledge has been developed around this topic. David Gefen (2000, 2003) has contributed significantly in this area and identifies trust as a combination of: (1) beliefs about integrity, benevolence of another party; (2) general belief that another can be trusted; (3) feelings of confidence and security in the response of another; or (4) a combination of these various characteristics. Further, Corritore et al. (2003, p. 740) provide a definition of online trust that includes cognitive and emotional elements. They suggest that trust encompasses “an attitude of confident expectation in an online situation or risk that one’s vulnerabilities will not be exploited”. In e-collaboration, the user ideally seeks confidence in the computer mediated experience and that over time a trusting relationships with other users can be established.

There is a growing reservoir of research on the role of trust in computer-mediated communication (CMC) including virtual collaborations and long-distance communications in instant messaging, blogs, e-mail, teleconferencing, video conferencing, or virtual teams (Bos et al., 2002; Jarvenpaa et al., 1998; Walther, 1992). Trust research in these contexts has focused on interpersonal relationships as they pertain to collaborative work (Hossain & Wigand, 2004). More specifically, how initial trust develops is important to understanding how long distance relationships are developed and nurtured (McKnight et al., 1998). Taken collectively, it is expected that if Web based technologies are able to elicit trust in users, then communication as in GSS and other forms of e-collaboration will be enhanced.

Alternately, it has been proposed that CMC can potentially delay trust formation related to a slower rate at which clues about a partner can be gathered in a virtual setting (Walther, 1992). One might expect trust formation to be especially difficult in a CMC context that transgresses cultural boundaries. Not only is there difficulty in obtaining cues about a partner due to spatial distance, but
due to cultural distance or unfamiliarity as well. Little work has been undertaken on cross-cultural effects and online trust, and studies completed in this area have generally yielded inconclusive results. For instance, trust in e-commerce was examined between American and Taiwanese participants and systematic differences were found concerning privacy, but no differences were evident related to culture (Lui et al., 2004). In another study on culture and trust, differences existed between American, Canadians, and Germans with Japanese (Cyr et al., 2005).

In sum, trust formation is important in virtual relationship building just as it is in face-to-face encounters. In settings where visual cues are absent, and when cultural differences prevail, e-collaboration is likely to be more challenging. Once again, it is expected that cultural sensitivity to user values and expectations is a precursor to the development of successful working relationships using technology. In some instances face-to-face encounters prior to virtual collaboration may be useful both within and between cultures.

SATISFACTION AND E-COLLABORATION

For decades satisfaction has been important to the information systems research agenda, and has frequently been used as a facsimile for success. If users are satisfied, then they have positive attributions to the system, service, or user encounter. Numerous determinants of satisfaction have been proposed and confirmed including system quality and information quality (Doll & Torkzadeh, 1988). With the advent of the Internet many other studies have considered the user interface as a potential contributor to online satisfaction. This has included the “ambience associated with the site itself and how it functions” (Szymanski & Hise, 2000, p. 313).

Disconfirmation theory has applicability to e-collaboration. This theory stipulates that satisfaction is determined by a comparison between the perception of performance and a cognitive standard (Oliver & DeSarbo, 1988). In other words, does the object meet a user’s expectations? Marketing studies have a history of using disconfirmation theory to explain or predict satisfaction in the context of products or services. The theory has also been used to measure manager satisfaction with group decision support systems (DeSanctis & Gallupe, 1987). In this instance, user satisfaction is broadly defined as the “multidimensional attitude towards various aspects of MIS such as output quality, man-machine interface, EDP staff and services, and various user constructs such as feelings of participation and understanding” (Khalifa & Lui, 2004, p. 39).

In other work, Mahmood et al. (2000) compiled studies on user satisfaction over a ten year period and determined that satisfaction is mainly affected by user background, perceived benefits, and organizational support. User background is determined by user experience, user skills, and user involvement in the system development process. Perceived benefits are measured by user expectations, ease of use, and perceived usefulness of the system or object. Organizational support is driven by user attitudes toward information systems, organizational encouragement, and perceived attitude of senior management. Expectations about the system are thus related to disconfirmation theory and are shaped by personal experience and environmental factors (Khalifa & Lui, 2004). It is anticipated that the above determinants of satisfaction are tempered by culture, and the user experience of in-country system characteristics and support.

Taken collectively, it is feasible that culture factors into the satisfaction equation and is based on different values that prevail across cultural groups (Hofstede, 1980). As one example, Hofstede outlines different levels of uncertainty avoidance in diverse cultures, when members of a group seek to reduce personal risk. Such risk aversion occurs in traditional settings, but could
be extrapolated to online encounters as well. Also part of Hofstede's framework, certain cultures such as those in North America prefer to operate more independently, while Asian cultures tend to operate in groups and collectively. While these are cultural generalizations, in the context of e-collaboration there are indications that different expectations regarding online interactions are likely. If a discrepancy exists between cultural values and the collaborative format, then user satisfaction is less likely to occur.

Pulling together the threads of research as it affects various forms of e-collaboration, it appears that user values and expectations including those that are culturally based will influence perceptions of user satisfaction. User values, norms, and experience are likely to have an impact on reactions to technologies used for computer mediated communication.

CONCLUSION

The introduction and use of e-collaboration technologies has met with mixed levels of success (DeSanctis et al., 1993; Kock, 2004). In this article, it has been argued that culture is important to better understanding how e-collaboration operates. It is an area that has been largely overlooked to date. However culturally sensitive user interface design is described to potentially increase user trust and satisfaction, therefore enhancing the user e-collaborative experience.

In the future it is expected that designers of e-collaboration tools or interfaces will factor in cultural characteristics to the technology system. As one example there is growing evidence that culturally appropriate user interfaces are important for educational courseware (Pfremmer, 2004). Cultural acknowledgement can impact user impressions of usability, accessibility and acceptability of software and will influence factors such as ease of learning, efficiency of use, memory ability, error frequency and severity, and subjective satisfaction (Kondratova & Glodfär, 2005). Also related to e-learning, collaboration and learning can be hindered by teaching style differences, problems related to different educational values, or diverse language and semantics (McLoughlin, 1999).

Similarly, cultural characteristics should be considered as they relate to effective functioning of project management teams, group support systems, and other forms of e-collaboration. A challenge for researchers is how to best adapt various technologies for diverse cultures. In an area ripe for future enquiry, investigations might focus on how trust and satisfaction are enhanced in e-collaboration, and factors that influence this process.

REFERENCES


Enhancing E-Collaboration Through Culturally Appropriate User Interfaces


**KEY TERMS**

**Culture:** Patterns of thinking, feeling, and potential acting learned throughout a lifetime, and applicable to groups of people often from the same nation state.
Enhancing E-Collaboration Through Culturally Appropriate User Interfaces

**E-Collaboration:** Collaboration among individuals involved in a common task using electronic technologies.

**E-Learning:** Learning in groups or individually as facilitated by online interfaces.

**Online Satisfaction:** Positive attributions by a user toward an online system, service, or encounter.

**Online Trust:** A belief that an online user has confidence in a computer mediated experience.

**User Interface:** The use of technology to connect users in an online environment.

This work was previously published in Encyclopedia of E-Collaboration, edited by N. Kock, pp. 240-245, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 5.6
Cultural Barriers of Human–Computer Interaction

Deborah Sater Carstens
Florida Institute of Technology, USA

INTRODUCTION

Information and Communication Technology (ICT) researchers and practitioners are well aware of the cultural challenges brought on by a global market (Smith, 2004; Smith, Dunckley, French, Minocha, & Chang, 2004). However, there are many unresolved problems concerning the extent to which culture influences ICT usability. Businesses use ICT in the form of databases to house customer information, Web sites enabling customers to place orders, information systems for management or suppliers, training systems, and as products sold to customers. Internet growth enables businesses to expand their customer base to international markets. Thus, businesses benefit from the explosion of Internet usage but may be challenged by how to best meet the needs of their multi-cultural customers, suppliers, and employees. There is a need to develop a model of cultural barriers to human-computer interaction (HCI). With all of the technology in use today, along with the different cultures that interact with ICT, it is important to identify a model of ICT and the HCI barriers produced by it to better help designers of ICT avoid these technology pitfalls. Figure 1 displays how the incorporation of technology, people, and culture into businesses must be carefully positioned together to optimize the success of all involved.

This article examines cultural barriers to HCI and outlines a model to help designers of ICT avoid these barriers so as to enhance a company’s ability to conduct business internally and with international businesses and customers. The article addresses topics of interest to ICT practitioners and researchers alike. Current services available to businesses that support effective international HCI are discussed. Current research and future research opportunities in the field of international HCI in ICT are also examined.

BACKGROUND

It is important to define HCI as the study of human behavior in interacting with any computer-based device such as a Web page, information system,
or other technology. Through the study of HCI, researchers optimize interactions between the human and technology through better design of the ICT. Traditional ways of conducting business seldom exist in a society experiencing growth in the Internet, globalization of trade, and digital integration internal and external to a company. The role of ICT including e-commerce and information systems has become more complex as businesses continue expansions into global markets (Laudon & Laudon, 2004). With businesses having multi-cultural customers, suppliers, and employers, there are demands for businesses to utilize ICT that adequately meets the needs of cultural groups that interact with the technology. This section discusses how international HCI is an important aspect of ICT. In fact, there are businesses that specialize only in servicing other businesses in providing multi-cultural aspects to ICT. Furthermore, current research in the field of HCI is addressed to uncover cultural barriers to HCI that potentially produce ineffective ICT for specific user communities.

Today, HCI is in the spotlight as companies outsource software development or Web sites, operate e-commerce businesses, and extend their services to international customers. The globalization of business has pushed industry into finding ways to incorporate international HCI. Usability is a topic under the umbrella of HCI that is important to design into any ICT, especially systems utilized by a multitude of international users. Zwick and Dholakia (2004) suggest that adoption of any ICT brings about multiple issues to consider regarding social, political, economic, and cultural implications. ICT such as Web sites are generally adopted to solve a business need, but this article also suggests that with the introduction of technology, businesses may inherit new challenges before achieving full resolution of their business need. With the globalization of businesses, cultural barriers must be identified and
corrected in ICT-enabling companies to continue relationships with different cultural groups within their customers, employees, and suppliers.

There are ICT practitioners dedicated to servicing businesses to support effective international HCI. Current services available to support multi-cultural HCI business needs consist of companies that provide written translations, verbal interpretations, multilingual desktop publishing, audiovisuals, Web site localization, globalization consulting, language services, global branding support, content development, end-user experience designs, translation of business correspondence, international user research, usability testing, etc. Aplomb Translations (2004) is an example of a company that specializes in providing multi-cultural language services. Aplomb focuses on audio and video transcription, subtitling, voice-overs, and Web site translation. Another company, ABC Translations (2004), focuses on services such as Web site and software localization. Web site localization assists a company with services such as providing multilingual online marketing strategies to globalization-friendly sites. Software localization uses in-country specialists to create localized versions of software applications to serve specific cultural user communities. Crosscultural (2004) provides global online marketing, helping companies expand their Internet presence into different cultural communities. The expansion into global markets creates open doors for practitioners and researchers providing services and solutions that help businesses to better meet the demands of multi-cultural customers, suppliers, and employees through effective international HCI in ICT.

Research has been performed in the field of HCI and the subtopic of international HCI has increasing popularity with the globalization of business. Cultural barriers to verbal communication between individuals from even different English-speaking countries can be difficult (Norman, 1999). Barriers to international HCI are therefore concerned with not only language but also other factors as well. Heldal, Sjøvold, and Heldal (2004) suggest that a user’s perception of a Web site will vary from culture to culture and is influenced by every aspect of a site, even beyond content and language. This experiment focused on evaluating Web sites using subjects from Norway and South-Europe. Aspects of a site such as site organization, frustration and innovation were rated differently by the Norwegian participants than the South-European participants. These findings suggest that differences in perceptions or impressions of a site can be related more to culture than the actual design of a site. An experiment conducted by Siala, O’Keefe, and Hone (2004) indicates that individual religious backgrounds may also affect online trust towards electronic commerce. They found that people and communities of people sharing common values and personality traits from different cultural groups can build in-group trust (Siala et al., 2004). The study compared individual responses based on their religious affiliation (Christian, Muslim, etc.), gender, and ethnicity. The study measured the extent to which religion and possibly in-group trust affects a participant’s buying habits and inclinations to buy online from within a religious group. Findings indicate that religious backgrounds did make a difference as some participants appeared to be more consumer religio-centric. Trust and attitude in a Web-based retailer varies within cultural groups and increases the importance of businesses identifying with different cultural groups where trust and attitude affect buyer behavior (Siala et al., 2004). Another study by Cannon and Perrault (1999) suggests that satisfaction is important in relationship marketing, which is also linked to communicating proper and distinctive impressions. There are also interrelated aspects of these cultural barriers, as Heldal et al. (2004) suggest that poor usability of a site may give a customer a bad impression of the company and its products. These studies serve to demonstrate that international HCI is important in the design of effective ICT.
The term “usability” refers to the capacity of software to easily and quickly perform the tasks that users seek (Dumas & Redish, 1999). Nielsen (1993) further classifies usability as either practical or social acceptability. For the purposes of this article practical acceptability examines the usefulness of the ICT being able to carry out needed user functions. For example, if a person from England uses the search feature on an American Web site to find the term “organisation,” the site should accommodate the language of the user even if it is not in the language of the designer of the site (i.e., Americans spell the term using a “z” instead of a “s” as in “organization”). Social acceptability refers to whether the technology is considered ethical within the purpose of the technology. This area is obviously complex and important since different cultures may have different impressions of ethics. Companies obviously should be interested in the effect of how pleasant a customer’s online shopping experience is on sales. Furthermore, it is important that employees have very usable ICT systems because employees using difficult systems may spend longer performing set tasks and this could also impact morale and subsequently productivity.

Wickens (1992) developed a human-information-processing model that addresses the internal and external information processing that a human goes through in performing an information task. The model is useful in the application of ICT design of HCI issues as culture impacts the perception and interpretation of information. For example, if the language conversion feature on an online purchasing site is worded in English instead of having pictorial language indicators (such as flags), many users will have difficulty in understanding the site. More recent work by Te’eni and Sani-Kuperberg (2004) performed research extending the human-information-processing theory into the identification of Levels of Abstraction (LoA). LoA refers to the process of users getting lost in the details of information search. When users shift focus to reading the online help feature, they are switching from one level of abstraction to another. LoA research is a detailed task analysis of a user operating technology. If LoA can be generalized for specific cultural communities, it may also be useful in designing ICT for multi-cultural information processing and thus extend previous information processing research.

CULTURAL BARRIERS TO HCI

ICT was earlier described to include any computer-based device. For businesses, common ICT consists of customer relationship management databases, computer-supported collaboration, management information systems, Web sites, training tools, supply chain management systems, and technology produced or used by a company. The literature asserts that cultural barriers to adoption and effective use do exist and greatly vary from culture to culture (Heldal et al., 2004; Siala et al., 2004; Smith, 2004; Smith et al., 2004; Zwick & Dholakia, 2004). The literature suggests that some of the common barriers that exist include poor organization, hard-to-use features (not user-friendly), frustration, confusion, non-usable features, trust, religion, ethics, and lack of satisfaction (Cannon & Perrault, 1999; Heldal et al., 2004; Nielsen, 1993; Norman, 1999; Siala et al., 2004). If Nielsen’s (1993) earlier work is considered, new barriers not previously identified include religion, trust, and ethics. Figure 2 displays a model that shows the impact that cultural barriers have on ICT for a business. The negative consequences of the barriers that impact customers, businesses, employees, and suppliers are extracted from the literature and include poor impressions of a business, decline in sales, and company efficiency and productivity (Lazar, 2001; Heldal et al., 2004; Cannon & Perrault, 1999; Siala et al., 2004). In a world where businesses are under constant pressure and competition to attract customers, to maintain dedicated employees, and to develop innovative products, it is challenging for companies...
to devote the necessary time and attention that HCI demands in meeting multi-cultural needs to adequately optimize businesses, technology, and people. The model presented in Figure 2 aids designers of ICT in considering cultural barriers to HCI and its impact on businesses, employees, customers, and suppliers.

**FUTURE TRENDS**

Future research opportunities for ICT practitioners and researchers are discussed in this section. The current implications are that ICT will continue to grow and eventually pulse through every aspect of an individual’s environment and the emerging field is in the social appropriation of ICT. The social appropriation of ICT brings with it an increasing need to view adoption from a users perspective in an environment of almost infinite choice. In such circumstances HCI will continue to be a significant part of the new and changing adaptations of ICT.

Obviously this will require increased effort on behalf of researchers and designers to define and mitigate barriers to ICT usability. Culture has been identified as a significant issue in the increasingly globalized world and follow-on research is required to validate the cultural barriers to the HCI model proposed. The next steps include the identification of guidelines that could be developed to design out cultural barriers to HCI addressing how to design ICT that is considered easy to use.

Luckin (2003) identified a checklist to help learners use multimedia. The checklist includes identification of the goal for using the technology, reminders of the goal, guide to sub-goals, and a model answer, thus providing feedback to the user for future use of the media. Given the increasing importance of social appropriation, research should be expanded to identify a checklist for use by designers of ICT to design out cultural barriers to HCI.

There is a need for research addressing the influence of animations on interaction with Web
Cultural Barriers of Human-Computer Interaction

pages to measure the effectiveness of these interface enhancements (Schaik and Ling, 2004; Lee & Benbasat, 2003). This also presents an opportunity to identify the cultural preferences of such features. As Intranets also utilize such animations, research could be conducted to determine if productivity of employees is enhanced when such features are active. Furthermore, design teams could be studied to determine the successful incorporation of international HCI in ICT. Therefore, the effectiveness of multi-cultural design teams should be studied. Several researchers have explored international usability testing and have found challenges in the cost of operation across cultures as well as difficulty in communications between the usability professionals and users of different cultures (Dray, 2001; Murphy, 2001; Yeo, 2000, 2001).

As the world continues to push technology into global markets, a gap analysis effort could be performed to identify current and potential future needs of businesses versus current available services in terms of effective HCI in ICT.

CONCLUSION

The world has been revolutionized by the abundance of ICT that greatly transforms the way businesses and individuals communicate and operate. International HCI looks at ways that information displayed through ICT can be enhanced to optimize use in multi-cultural environments. While there is increasing recognition of the importance of the impact of culture on effective use of ICT in a rapidly globalized world, there is not a substantial body of research or practice that matches the expanding use of ICT. As ICT favours increasing economies of scale, business (both large and small) will drive the demand for answers to the impacts of culture on the social appropriation of ICT. This is of vital importance to regional business as it seeks to build interactions with particular regions across the world.

REFERENCES


Cultural Barriers of Human-Computer Interaction


**KEY TERMS**

**Barriers to Human-Computer Interaction:** Anything that poses as a challenge to a human interacting with technology. Examples include poor organization of a Web site and language conversion issues.

**Consumer Religio-Centrism:** Individuals who are so strongly committed to their specific religious group (Christian, Muslim, etc.) that their buying preferences consist of purchases from companies owned or operated by individuals with their same religious beliefs.

**Customer Relationship Management:** Technology in a business that houses information on customers. The information systems give businesses an integrated view of each customer that may consist of demographic data, buying trends, and financial data.

**Enterprise Functions:** Enterprise functions are composed of typical functions of a business such as human resources, marketing and sales, manufacturing, accounting and finance.

**Human-Computer Interaction (HCI):** Study of human behavior in interacting with any computer-based device. HCI is concerned with
identifying ways to optimize such as through the design of technology and the relationship between humans and computers.

**In-Group Trust:** Communities of people that have similar beliefs, values, and personality traits induced by cultural forces and are more willing to identify members inside their group as trustworthy. Individuals within these communities obtain role models and opinion leaders within their group that influence individual’s beliefs and behavior to include buying decision processes.

**Information Processing:** Study of how humans interact within their environment while interpreting information to identify a decision. Humans are a unique form of a machine and must process information in order to formulate a decision.

**Management Information Systems:** Systems utilized by management-level employees to seek information pertaining to the operations of their business. The information gives management insight into potential market or financial trends as well as the productivity and efficiency of their employees and equipment.

**Supply Chain Management:** Automates the relationship between businesses and their suppliers to optimize shared processes. An example of this relationship is suppliers through the Internet viewing supply levels of a company and responding automatically through sending more supplies without the company directing the supplier to do so.

**Usability:** Study of the usefulness of technology as it relates to an individual being able to successfully perform their task with ease. Usability is measured through the ability of the technology to do the right job and to do the job right.

*This work was previously published in Encyclopedia of Developing Regional Communities with Information and Communication Technology, edited by S. Marshall, W. Taylor, and X. Yu, pp. 146-151, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 5.7
Technology and Culture: Indian Experiences

Ramesh C. Sharma
Indira Gandhi National Open University, India

Sanjaya Mishra
Indira Gandhi National Open University, India

ABSTRACT
This chapter discusses the deployment of e-learning technologies in the context of how they are helping towards preserving and disseminating knowledge on Indian cultural heritage. An analysis has also been offered as regards how the technologies like e-learning initiatives have their impact on sociocultural settings within Indian context. This chapter attempts to understand and frame Indian culture and experiences through ICT and e-learning practices, and how the differentiated learning needs of multicultural society can be addressed.

INTRODUCTION
Culture has been important, perhaps one of the most important concepts, of anthropology. A surviving culture is always dynamic in nature; it evolves over time and, at the same time, maintains its identity. Matsumoto (1996, p. 16) defined culture as, “the set of attitudes, values, beliefs, and behaviours shared by a group of people, but different for each individual, communicated from one generation to the next.” It is these values and beliefs which make a culture that can survive against all odds.

Indian civilization is one of the oldest civilizations of the world (dates back to more than 5,000 years). In spite of various foreign invasions and dominations, visits by scholars in search of knowledge, India’s cultural own identity has been maintained. It has adopted the good from them while rejecting those that might have destroyed its unique unity in diversity feature. This nation has witnessed many social, economic, political, and technological changes. Here, science and spirituality both exist. Indian architecture, festivals, dance forms, painting, sculptures, cuisine, literature, and teacher-taught relations, all have
different dimensions of extraordinary variety. This cultural heritage has been passed onto from generation to generation by appropriately preserving, promoting, and disseminating modes be it oral, written, or now electronic.

INFORMATION COMMUNICATION TECHNOLOGY (ICT) IN INDIA

The government of India has accorded high priority to the deployment of ICT for social and community development. Since early 1970s, India has witnessed constant growth in the area of telephone density, Internet penetration, establishment of radio and television stations, broadband connections, and affordable computers and peripherals. These have become within easy reach of the educational institutions, businesses, and individuals and so forth.

Starting with nearly a thousand Internet users in 1992, over 5 million users were enjoying Internet in 2000 (GOI, 2003). TRAI (Telecom Regulatory Authority of India, 2005) reported that there were 105,000 Internet cafes in India; telephone density was 0.6% in 1991, which increased to 11.7% in 2006.

Telephone, radio, and television has been widely used in educational settings in India (Sharma, 2002; Vyas, Sharma, & Kumar, 2002a, 2002b). One-way-video two-way-audio teleconferencing is quite effective for content presentation, learner-expert interactions, and learner-supported activities (Mishra, 1999). In addition satellites have also been considered to be useful in catering to the educational needs of the society. For example, EDUSAT is the first Indian satellite designed and developed exclusively to meet the demand for an interactive satellite-based distance education system for the country (Sharma, 2006). All these different instruments of ICT, like radio, television, teleconference, satellite, Internet, have contributed to the community development under various schemes of the government.

ICT FOR CULTURAL HERITAGE

Different measures for the preservation, transmission, and advancement of languages and culture have been adopted by the communities. With the physical expansion of the world, different cultures and languages have realized the increasing importance of having dynamic and vibrant mechanisms that can help them maintain their identity and foster progress in a multicultural learning environment. Odasz (n.d.) states, “The world’s diverse cultures jointly represent the full cultural genome of humankind’s search for individual and group identity and meaning” and exert pressure to record this important “shared story of humankind.” The sooner actions are taken to save the cultural knowledge of our ancestors, the better will be, as it is feared that nearly half of the world’s 6,000 languages may disappear in one lifetime. Odasz (n.d.) recommends, “The vast cultural knowledge of our elders must be recorded via multimedia storytelling for preservation while they (our elders) are still with us.”

The use of ICT to the cause of culture has been best exemplified through different kinds of technological tools (radio, television, Internet, etc.) being heavily used by Indian communities to maintain or to create new relations. The social networking tools, like MySpace (http://www.myspace.com); Ning (http://www.ning.com); Second Life (http://www.secondlife.com) or Orkut, are connecting people across different cultural and social backgrounds. Cellular phones have become an integral part of common people, due to slashing of rates prompted by the entry of many players and due to the Government of India’s intention to reach to the masses through telephone network. SMS poll is nowadays a common phenomenon in case of any national event where the viewpoint of masses has a say. This is becoming a cultural bonding technology even in some cultural events when the contestants request the people to vote for them through SMS like reality shows on television. Cole and Crawford (2007) consider ICT tools as
an effective means of enhancing communication among communities, building peace and facilitating better understanding of each other. A culture is what we think, what understanding we have for others, and how we behave. Manjul (2005) reported an "alternate media," in the form of a rural newsletter that was initially launched as a communication channel among the women themselves, but by the impact it had on the sociocultural issues, it evolved into a mechanism whereby it addresses the community as a whole, and a great tool for women empowerment.

TECHNOLOGICAL INITIATIVES AND CULTURE

ICT applications in India have been used for educational, economic, and social development of Scheduled Castes, other Backward Classes, and minorities, as well as protection of rights and opportunities for the welfare of disabled persons, children in distress, victims of drug addiction, and the senior citizens. Establishing a telephone help line for disabled persons in Mumbai has been a significant achievement. Similarly, a toll-free telephone service, “Childline,” provides help to children in distress or by adults on behalf of such children. This facility was significantly beneficial to Tsunami affected areas during 2005, like Kanyakumari, Nagapattinam, Kanchipuram, Cuddalore, Kollam, and Port Blair and so forth.

Although many religious communities in India possess strong, gender-related norms, situation is comparatively serious in Muslim community. Girls are not encouraged to leave home without a veil, education is low priority, early child marriage is common. Sharma and Maindiratta (2005) reported an interesting case, where the effort by Datamation Foundation Trust to establish an ICT centre at Babool-Uloom Madrasa (a religious residential school providing learning opportunities to boys from poor families) in a densely populated and low income area of New Delhi initially met with cold response. The Trust wanted to set an ICT centre to address urban poverty and increase women empowerment, and the project was initiated, keeping in view local cultural values. Internet was also used (using eNRICH, a local Web-based browser) to encourage local talent and cultural heritage. Cross-cultural experiences were made available to them to learn about the communities in other countries. Slowly, the results showed their effects and the centre started gaining popularity. This was most conspicuous during the month of Ramadan, when the center was to be closed for the month. The girls convinced the concerned authorities successfully to let the centre remain open for some time. That was unimaginable before, as the girls are not allowed to have such a dialogue. This clearly establishes the impact of the ICT on the cultural values.

Governmental Agencies in India and Projects

In India, technological mechanisms are being used for maintenance and conservation of the country’s heritage, ancient monuments, and historic sites; promotion of literary, visual, and performing arts; administration of libraries, museums, and institutions of anthropology; maintenance, preservation, and conservation of archival records and archival libraries; research and development in the conservation of cultural property; and promotion of institutional and individual initiatives in the field of art and culture. Technology is helping out in generating cultural awareness at the grassroots level to promote cultural exchanges at an international level. Detailed information is also made available on Indian history, literature, philosophy, religions, Sufism, art, architecture, astronomy, astrology, mathematics, medicines, and physical sciences.

Some of the important organizations and agencies in India dealing with preservation, conservation, and spreading awareness of culture are:
Cultural Informatics Laboratory

Cultural Informatics Laboratory (CIL) establishes a synergy between the disciplines of art and information technology leading to usage, development, and demonstration of new technology and cultural documentation.

Kalasampada

The project is sponsored by Ministry of Communications and Information Technology, Govt. of India, (MCIT) under digital library initiative (DLI) – India. The prime focus of this project is to develop a databank of cultural heritage housing over 4.5 million pages of manuscripts, 4,000 photographs, and 200 hours of audio and video.

Content Creation and IT Localisation – Network (COIL-NET)

Sponsored by the Ministry of Communications and Information Technology, Govt. of India, a Web site on “Cultural Heritage Digital Library” has been developed in Hindi with special focus on Hindi Speaking region mainly states of Uttar Pradesh, Uttarakhand, Madhya Pradesh, Chhattisgarh, Bihar, Jharkhand, and Rajasthan.

Manuscripts Digitization Projects

The digitization of rare manuscripts written in archaic languages and scripts such as Arabic, Persian, Sanskrit, Hindi, Urdu, Turkish, and Pashto, and so forth, is an important step towards preservation of Indian cultural heritage.

Ajanta - In-house CD-ROM Project

This CD-ROM project provides wide-ranging knowledge and visual experience (in the form of 1,500 illustrated images, articles from eminent authorities, Bibliography, Glossary, etc.) on Ajanta, a major heritage site of India, included in the UNESCO’s list of World Heritage monuments.

Eternal Gandhi – Multimedia Exhibition

The Eternal Gandhi is state-of-the-art digital multimedia exhibition, which deals with information on the historical events of the life and message of Mahatma Gandhi. The exhibition also presents a wide spectrum of information technology visions inspired by the thought and teachings of Mahatma Gandhi.

National Multimedia Resource Centre (NMRC) project

The National Multimedia Resource Centre (NMRC) project (launched in 1998 and sponsored by the Department of Information Technology, Ministry of Communications & Information Technology, Government of India) identifies and disseminates cost-effective applications of multimedia among the masses.
Indian Heritage Portal

The Indian Heritage Portal developed by C-DAC (Centre for the Development of Advanced Computing) provides information on Indian scriptures like the Vedas, Bhagavad-Gita, Mahabharata, and Dhyaneswari and so forth.

E-LEARNING FOR CULTURAL ISSUES

E-learning is increasingly considered as a means of greater access to education and development (Mishra & Sharma, 2005). Such projects have their impact on rural, illiterate, or women populations, the upliftment of whom is of prime importance. These have greater impact on the cultural issues also. E-learning practices have been found to be an important tool in preserving, promoting, and advancing cultural heritage. The initiatives taken by Kerala Government through its e-tourism have generated a world wide interest. Another successful example is Tamil Virtual University (http://www.tamilvu.org/) established as an Internet-based resource that provides details on the history, art, literature, and culture of the Tamils. This project is very popular among those Tamils who have settled in various parts of world, and interested in knowing about the cultural heritage of Tamil community. Indian Tobacco Company (ITC) is one of India’s leading private companies, and has initiated an e-Choupal effort (http://www.echoupal.com/) that places computers with Internet access in rural farming villages (http://www.digitaldividend.org/case/case_echoupal.htm ). A “Choupal” is a local village gathering place where the villagers gather and discuss their problems or issues.

Culture and Language Instructional Context

Many cultures (like rural folk in Gujarat and Rajasthan State in India) use narration and storytelling as a way of passing on the information from past to the present generation. Narration enables the learners to participate in the process and becomes the core of experiential and contextual approach to learning.

FUTURE TRENDS

Different technologies have different characteristics and thus, impact differently on the cultures (Idrus & Sharma, 2007). A radio or a television may impact differently than Internet. It depends on the nature of technology, that is, be it one-way or two-way/synchronous or asynchronous, how the cultural issues are going to be affected by it. It would be significant to understand in what way innovative ICT approaches can maximize the impact of different technologies by integrating them (Girard, 2005; Mathison, 2005). For example, in the Kothmale Community Radio Internet Project in SriLanka (http://www.kothmale.org) Internet is used to search the information and radio to disseminate it. Also, the listeners can send the questions to the radio station either by post or calling through telephone. Then the team at Radio Station uses Internet to search for needed information, translates it in local language, and broadcasts the answer.

Further, the research on measuring the impact of ICT on community development is still in its infancy in India. There is a strong need to undertake microdata analysis to identify the extent to which ICT changes the performance of small and medium enterprises in developing countries.
FUTURE RESEARCH DIRECTIONS

Future research directions must address the issues like how do the technology programmes deal with different communities and different cultures and languages? Which interventions need to be customized? To what extent is a programme effective in a specific community? Is there any supportive culture for using ICT for bringing out community development?

CONCLUSION

India is a vast country covering wide geographic area, divided into numerous cultures and languages. Over 700 million Indians live in rural areas. Of these, around 190 million live below poverty line. India has perhaps the largest number of poor people in the world. The situation further aggravates, as 84% of our illiterates live in villages. The diversity and magnitude of these problems create huge challenges for the ICT channels, which appear to be disorganized and with little effectiveness for communication and resource sharing. We need to have an in-built feedback mechanism and supporting research studies to showcase the indigenization of culture through ICT.

Culture of a place is reflected through group behaviour, social norms, how the group thinks, behaves, reacts, and responds to a situation. ICT can be an effective tool by encouraging cross-cultural exchanges, while at the same time bringing out positive social changes, by developing mutual respect for each other's culture, sharing of opinion. Proper networking of different agencies in a planned way would tackle sociocultural issues. The public-private partnership involving government would create innovative solution for society’s need. The invention of the simputer in India is an example.

Issues that may arise while developing culturally sensitive instructional material are values and attitudes, ethical perspectives diversity within groups, historical perspectives, socioeconomic perspectives, social roles, social networks, learner expectations, learning styles, and opportunities for interactions (Collis 1999; Dunn & Griggs, 1995; McLoughlin, 1999; Powell, 1997; Reeves, 1997and). Gjedde (2005, p. 104) suggested that issues, such as social and cultural backgrounds and gender-specific interests, must be considered when developing meaningful content and creating narrative e-learning environments. Singh (2003, p. 100) suggested steps to deal with problems faced during the development of e-learning in India viz. the diversified geographical spread; the use of different languages in different states and regions; the communication infrastructure; and the illiteracy. The language problem is being tackled on many fronts. The Microsoft XP desktop operating system has the ability to support nine Indian languages. Modular Infotech Pvt. Ltd., a Pune (Maharashtra)-based company, is actively working on the language front. It has partnered with Microsoft to develop various keyboard handlers for Indian languages.

The Government of India has urged upon private-sector companies to make a commitment to rural areas. India is looking forward to establishing a “Universal Service Fund” to service the rural IT plan. The Government of India has reduced telephone call rates to promote the use of the Internet and make the Internet and telephone services affordable to the general public. The Ministry of IT holds regular meetings to promote the concept of e-learning. The fast developments in information technology and products have revolutionized the ways of teaching and learning across different cultures. The content generated with IT applications acts as a vehicle for knowledge that is well-suited to many cultures and languages across the globe. The Internet tools in the form of Web-logs or pod-casting provide great opportunities for the preservation, expression, strengthening, and dissemination of cultural identities and languages (Mishra & Sharma, 2004).
REFERENCES


**ADDITIONAL READING**


ONLINE RESOURCES

• Association for Progressive Communications: http://www.apc.org/
• Bytes for All: http://www.bytesforall.org/
• Content Creation and IT Localisation – Network (COIL-NET) (http://www.tdil.mit.gov.in/coilnet/ignca/welcome.html)
• Culture and development. Unesco: http://www.unesco.org/culture/development/briefings/html_eng/
• Doors Delhi Web site http://doors8delhi.doorsofperception.com/
• Education Development Center: http://main.edc.org
• Global knowledge partnership: http://www.globalknowledge.org/
• Indian Heritage Portal http://www.cdac.in/html/ihportal/index.asp
• National Multimedia Resource Centre (NMRC) http://www.cdac.in/html/nmrc/about.asp
• OneWorld: http://www.oneworld.net/
• Sarai, The New Media Initiative Delhi India: http://www.sarai.net
• The Center for Knowledge Society: http://www.ict4d.info
• Worldbank: http://www.worldbank.org/

This work was previously published in Social Information Technology: Connecting Society and Cultural Issues, edited by T. Kidd; I. Chen, pp 181-189, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 5.8  
Intercultural  
Computer–Mediated  
Communication Between  
Chinese and U.S. College  
Students

Yun Xia  
Rider University, USA

ABSTRACT

This study incorporates the effects of culture and computer-mediated communication (CMC) in the investigation of Chinese college students’ use of English in communication with U. S. college students. Ethnography of communication was used as the method to uncover four language patterns in Chinese college students’ use of English: others-oriented talk, mentor-mentee talk with limited relationship development, adoption of others’ talk, and icebreaker talk. Chinese cultural values showed a strong effect on Chinese college students’ use of English in CMC. Chinese cultural values included Chinese significance of personal relationships, Chinese collectivism, Chinese understanding of authority, and Chinese use of CMC. The characteristics of CMC showed minimal effect as either a constraint or a way of encouragement in the four language patterns.

INTRODUCTION

Although English is a global language in intercultural computer-mediated communication (CMC), patterns, rules, and premises of English use vary across different cultures. Decades ago, Hymes (1968) proposed assumptions regarding speech and language in a cultural group. First, speech and language of a group constitutes a system. Second, speech and language vary cross-culturally in function. Third, speech and language activity of a community should be the primary object of attention. Thus, the study of particular use of English language in a cultural group in CMC should be
the focus that can uncover the resources of different communication contexts. The interpretation of meanings of those resources should provide guidelines for intercultural CMC in return. In this chapter, ethnography of communication was used as the research method to describe, analyze, and interpret Chinese college students’ English language activities and Chinese cultural resources regarding language activities that constructed a distinctive intercultural communication between Chinese and U.S. college students.

BACKGROUND
CMC and Intercultural Communication

In the literature of intercultural computer-mediated communication, there are two major perspectives. On the one hand, the focus on communicative effects of CMC argues that CMC with its unique characteristics can facilitate and contribute to intercultural communication. On the other hand, the evaluation of cultural factors in communication leads to the conclusion that communicative styles in CMC as well as attitudes and perceptions toward CMC are subject to cultural effects. However, few studies in the literature have incorporated both CMC characteristics and cultural effects in the examination of intercultural CMC. The current study adopts both perspectives in the evaluation of Chinese cultural effects and CMC’s effects on Chinese college students’ English communication with U.S. college students.

CMC includes different formats: e-mail, chatgroups, virtual worlds, and the World Wide Web. Across different formats, CMC is primarily in the form of written language. Compared with face-to-face communication, the written communication in CMC is “leaner” because the social context cues in face-to-face communication are missing (Kiesler & Sproull, 1992; Walther & Parks, 2002). The missing social context cues include people’s race, gender, social class, and accent (Sproull & Kiesler, 1991). The direct communicative effect of this characteristic is CMC’s potential for fostering flagrant and hostile language known as “flaming” (O’Sullivan & Flanagin, 2003). CMC is also ranked as low in media richness with less immediate feedback and limited personalization (Daft & Lengel, 1986; Trevino, Lengel, & Daft, 1987). Media with low richness communicate low social presence. In CMC, users may pay less attention to others’ presence. As a result, CMC makes communication more task oriented (Walther & Burgoon, 1992). Empirical studies (Jonassen & Kwon, 2001; Mazur, 2004) report that during group problem-solving activities, CMC participants produce fewer, but more task-oriented, messages than do face-to-face communication participants.

In intercultural communication, these characteristics of CMC encourage more participation from people who may normally feel intimidated in face-to-face interaction. As Warschauer and De Florio-Hansen (2003) stated, people in CMC do not know whether others are a dog, Black or White, male or female, or rich or poor. Thus, people who are less outspoken in face-to-face situations may contribute more in CMC with anonymity or less exposure (Simons, 1998; Tella & Mononen-Aaltonen, 1998). In some cultures, people take a passive rather than an active role in interaction. With low social presence, CMC’s task-oriented communication characteristic offers people an opportunity to make more contributions (Warschauer, 2000; Warschauer & De Florio-Hansen, 2003). CMC’s absence of nonverbal cues such as frowning and hesitating also makes interactions less intimidating and thus allows people in cultures that are less dominant to play a more active role in intercultural communication (Warschauer, 2000; Warschauer & De Florio-Hansen, 2003).

On the other hand, CMC as well as other media is subject to cultural influences (Leonardi, 2003;
Studies show that the differences between individualistic and collectivistic cultures affect the levels of satisfaction with e-mail (Chung, 1992; Scollon & Scollon, 1995; Simons, 1998). With social relationships constructed in language, an individualistic culture allows people to be more satisfied with task-oriented, short, and straightforward e-mail that has minimal context cues. With special forms of discourse that carefully preserve the boundaries between those who are inside members of the group and all others who are not members of the group, a collectivistic culture may require more context information for a successful communication in e-mail (Chen, 1998; Gunawardena et al., 2001).

The differences between high- and low-context culture also affect the use of CMC as well as other communication technologies (Chen, 1998; Chung, 1992; Leonardi, 2003; Xia, 2006). A high-context culture has most of the information in communication implicit (Hall, 1976). The meanings are to be inferred from the context or internalized in communicators themselves. A low-context culture has information explicit in the transmission of verbal messages. In a low-context culture, people tend to use verbal expressions explicitly, value self-expression and eloquent speech, and express their opinions directly. Studies (Chen, 1998; Chung, 1992) suggest that high-context cultures will be less satisfied with relationship building and feelings expressed in CMC. Low-context cultures, on the other hand, are likely to find CMC useful for expressing specifics in a short and straightforward style.

The differences between monochronic and polychronic culture also impact intercultural communication in CMC (O’Dowd, 2001). Monochronic cultures, such as the U.S. and northern European countries, are task oriented with emphasis on scheduling activities according to deadlines. Polychronic cultures, such as countries from the Mediterranean and the Middle East, prefer to combine activities and consider deadlines as objectives rather than as anything certain (Taylor, 2000). Communication between the two cultures can break down if people in a polychronic culture do not answer e-mails on time or do not work on a timetable for a synchronous CMC interaction (O’Dowd, 2001). Meanwhile, people from a monochronic culture may appear to be demanding to polychronic people when they try to get things done as fast as possible (O’Dowd, 2001).

**Language Analysis in Intercultural CMC**

Because CMC is primarily text based, the analysis of language use becomes critical to understand the nature of CMC and the relationship between language and social practice in a culture (Herring, 2001). However, the literature is particularly sparse on how particular language behaviors function in intercultural CMC. Only a few studies have explored unique cultural experience through language interaction in CMC. O’Dowd (2003) investigated intercultural experience in CMC and identified key characteristics of e-mail exchange that helped develop intercultural communicative competence. Choi and Danowski (2002) studied power-play negotiations of dominant and minority cultures in CMC through the analysis of structural patterns of online communication flow. In a study of how linguistic interaction patterns changed over time among a geographically and ethnically diverse group of young people, Cassell and Tversky (2005) uncovered the ways in which people from different cultural backgrounds constituted a community, spoke in a collective voice, and formed a unique culture. Through the analysis of written reports and interview notes, Ma (1996) analyzed the recurring themes in CMC between East Asian and U.S. students. After proposing and testing five propositions about intercultural CMC, Ma found that both East Asian and U.S. students were more direct in CMC than in face-to-face communication.
In the study of culture through language analysis in CMC, Herring and her colleagues played a significant role. According to Herring (1996), individual electronic messages are internally organized text, and texts’ distinctive schematic organization can be uncovered through language analysis. After studying one female-dominated and one male-dominated listserv, Herring found that messages posted by women were more interactive and contained more information while men tended to express their views critically. With the discovery of distinctive schematic text organization, Panyametheekul and Herring (2003) studied language use and culture in turn-taking of a Thai chat room. They found that Thai females had more frequent participation than males. This finding broke cultural expectations about the roles of women in Thai society and the role of women online.

From the intersection of the reviewed bodies of literature, the following are three research questions about Chinese college students’ use of English in intercultural CMC with the U.S. college students. With the three research questions, the study attempts to examine the effect of Chinese culture and characteristics of CMC on Chinese students’ use of English in CMC.

**RQ1:** What are the distinctive language patterns in Chinese college students’ use of English in intercultural CMC with U.S. college students?

**RQ2:** How does Chinese culture affect Chinese college students’ use of English in intercultural CMC with U.S. college students?

**RQ3:** How do the characteristics of CMC affect Chinese college students’ use of English in intercultural CMC with U.S. college students?

**MAIN FOCUS OF THE CHAPTER**

**Method**

In the study of this chapter, ethnography of communication was used as the research method. Ethnography of communication fills the gap between ethnography and linguistics. There is a widespread relationship between a language and a culture. The actual uses and patterns of use of language in different cultural groups need ethnographic description. The focus of ethnography of communication is on the way that communication in a cultural group is patterned and organized as systems of communicative events. An ethnographer of communication is “a naturalist, who watches, listens, and records communicative conduct, in its natural setting” (Philipsen, 1992, p. 7). With non-participant observation of intercultural CMC scripts and in-depth interviews in ethnography of communication, the researcher is able to describe, analyze, and interpret Chinese college students’ language activities that construct distinctive communication patterns.

The international exchange program in a private university of the U.S. was used as the research field. In this exchange program, a group of Chinese students come to the university and study business administration for a year each spring. In the fall before they come to the U.S., the university recruits U.S. students as Chinese students’ mentors. Still in China, each Chinese student communicates with his or her mentor through e-mail and other formats of CMC about culture, classes, campus housing, food, and weather. In this study, a group of 12 Chinese students was solicited as the participants. The participants were contacted and asked to keep their CMC scripts. After the participants arrived in the U.S., CMC scripts were collected and in-depth interviews were conducted. The researcher audiotaped all interviews, and all tapes were transcribed. All interviews were conducted in English.
In the analysis, elements of Carbaugh’s study (1989) about different cultural terms for talk were used. The messages from the terms for talk could be about communication, sociality, and personhood. Those elements enabled the researcher to start with the metacommunicative vocabulary in the interview transcripts that might indicate the way the participants chose to communicate with their U.S. mentors and about how they related to the U.S. mentors. With the metacommunicative vocabulary as the guideline, the researcher read and reread the transcripts of all recorded tapes and collected CMC scripts. During reading, patterns in the participants’ language behaviors were identified. After that, the transcripts and scripts were reassembled and reorganized according to the patterns. Examples and stories in Chinese participants’ communication were used to explain how their language behaviors were affected by Chinese culture and CMC characteristics.

Findings

In the analysis, four patterns that constituted Chinese participants’ English language use in CMC with their U.S. mentors emerged: others-oriented talk, mentor-mentee talk with limited relationship development, adoption of others’ talk, and icebreaker talk. For each of the four patterns, exemplars were used to illustrate Chinese participants’ use of English language. Some uses were experienced by individuals, but others were more universal.

Others-Oriented Talk

In Chinese participants’ communication, there was a clear pattern that showed much consideration of others. Some participants termed this pattern as “others-oriented talk.” The “others-oriented talk” was illustrated by Chinese participants’ extreme politeness, consideration of others as the priority, and talk about others in communication.

Although Chinese participants were more direct in CMC than in face-to-face communication, Chinese participants showed extreme politeness in CMC. The politeness markers, such as “please,” “thank you very much,” and “I appreciate,” appeared in Chinese participants’ e-mails and texts of instant messaging all the time. For example, one Chinese participant asked about the weather in the U.S.: “If it is not too much trouble, would you please tell me whether it is very cold in winter or very hot in summer there.” Sometimes, the U.S. students were even annoyed by the extreme politeness. One Chinese participant spent two paragraphs explaining why she did not e-mail back for two weeks. Her U.S. mentor started the response: “It is OK. You are not ignoring me. It is a waste of time explaining too much. Everyone is busy. I understand. Let’s talk about some serious stuff.” In the interview, Chinese participants explained that being polite showed their consideration of others because they were seeking help from their U.S. mentors.

Others-oriented talk was also about how Chinese participants talked about others rather than themselves. The talk about others could be a perfect example of Ma’s (1996) finding about East Asians’ indirect online communication that directed around another person other than himself or herself. Most Chinese participants loved to talk about their families, friends, and classmates. One Chinese participant mentioned in e-mail that she would bring a laptop to access the Internet. She could keep close contact with her parents and friends. With a computer camera, she could even see her parents as if she was around them everyday. On the other hand, most U.S students engaged in individual talk that was featured by personal plans and personal interests. One U.S. mentor talked about her trip to Puerto Rico and the Virgin Islands in the summer. She also talked about her love of dancing: “Dancing is my life. I’m actually the vice president of the dance team at the university and choreographer. So I like to go out...
to clubs and also we perform also at Basketball games and things like that.”

Others-oriented talk was directly related to the consideration of others as the priority in communication. Chinese participants tended to yield convenience to others and incorporate others’ feelings in CMC. One participant showed her consideration of others in the e-mail: “As your holiday break is coming, you must be extremely busy with your courses these days. I will be embarrassed to bother you with my questions. I will wait until you have all the tests pass.” After several weeks of e-mail communication, some of the participants started to talk about using synchronous CMC, such as MSN or AOL instant messaging systems. The major problem of using synchronous CMC was the time difference. Chinese participants always took the first step to offer the adjustment. One participant said in the e-mail: “Well, since you are busy with class work, I can stay late until 11 o’clock or midnight. It will be your 10 or 11 o’clock in the morning. It is easier for you.”

In the interview, Chinese participants used Chinese culture about the significance of others in a person’s life to explain others-oriented talk. In one participant’s words, others were an integral part of personal life: “I don’t even realize I talk about others a lot in the e-mails to my mentor. Well, for me, without others I will be like a leaf without branches and trunks. I will dry and die.”

In Chinese culture, the significance of others is closely related to Chinese values about smooth and harmonious personal relationships (Yang, 1994). Personal relationships are often encountered when seeking to establish friendships or even business partnerships in China. The way to establish personal relationships is to develop networks of mutual dependence. Personal relationships depend on the creation of obligation and indebtedness to each other. Thus, all participants thought that their questions and queries might add “extra burden” for their U.S. mentors. They all felt “indebted” to others whenever they sent questions to their mentors.

As for the effect of CMC characteristics on others-oriented talk, Chinese participants regarded e-mail as well as other formats of CMC as a constraint for their indirect way of communication. In one participant’s words, e-mail was “direct and formal” communication. E-mail did not work well as a way of communication for Chinese indirect communication. Since Chinese participants still engaged in an indirect way of communication, the characteristics of CMC as a constraint made Chinese participants uncomfortable. Interestingly enough, one participant talked about his changed attitude toward e-mail after coming to the U.S.: “I feel much more comfortable to use e-mail here. I think I am adopting the U.S. direct, straightforward way of communication. E-mail offers me the way to engage in direct communication.”

**Mentor-Mentee Talk with Limited Relationship Development**

Chinese participants used the term “mentor-mentee talk” to summarize the intercultural CMC between them and their U.S. mentors. Chinese participants always remembered their role as a mentee and the U.S. mentors’ role as a mentor. The role as a mentee defined what a Chinese participant could talk about and in what way a Chinese participant could talk.

First, mentor-mentee talk was illustrated by the topics that Chinese participants chose in CMC. In summary, there were six categories of topics that Chinese participants used: weather, schoolwork, textbooks, family, clothes, and campus housing. These topics were “safe for a mentee.” In the interview, Chinese participants explained that they never met their mentors and all they could talk about were the “general stuff in life they need to know.” Second, mentor-mentee talk was also about the way that Chinese participants chose to communicate with their mentors. Chinese participants’ extreme politeness also emerged in mentor-mentee talk. For example, one Chinese participant expressed his appreciation for the
answer about campus housing: “It is very kind of you to take such trouble to give me such a detailed explanation.” In the end of his e-mail, he expressed his appreciation again: “Please accept my gratefulness. Thank you very much.” In the interview, Chinese participants explained that they felt honored and grateful to have the U.S. students as their mentors. They should show great respect for their mentors in communication.

On the other hand, the U.S. students never treated Chinese participants as their mentees. They tried to make friends with Chinese participants. Thus, the U.S. students engaged in a different way of talk. Chinese participants termed their U.S. mentors’ talk as “friendship talk.” In their first e-mail, the U.S. mentors introduced themselves and always concluded with their willingness to be friends of Chinese participants: “Please feel free to contact me with any questions or concerns you may have about coming to the U.S. When you arrive in January, I will be ready and waiting to give you a tour of campus and help you get acquainted with the U.S. Think of me as a friend, you can approach me with anything.”

When one Chinese participant expressed the worry that she might bother her mentor too much, the U.S. mentor e-mailed back: “You are not bothering me. Do not worry. We are friends.”

In terms of the effect of CMC on mentor-mentee talk, Chinese participants had a limited relationship development that was encouraged by the characteristics of CMC. The limited relationship development between Chinese participants and their U.S. mentors was reflected in the way addressing each other evolved from formal to casual and no addressing at all. At the beginning, both Chinese participants and their U.S. mentors addressed each other very formally. One U.S. mentor initiated the contact in the first e-mail with “Dear Christina Yi Zhang” and concluded the e-mail with “Yours, Laurie Smith.” The Chinese participant also answered with “Dear Laurie Smith” and concluded with the term “sincerely” and her full name. After the second week, the U.S. mentor started to address her Chinese mentee with “Hey, Chris” and ended with only “L.” After the fifth week, the U.S. mentor used “Hey, hey” and sometimes started e-mails without addressing. The Chinese participant also switched to “hi” as the start and concluded with “c” in the end. In the interview, the Chinese participant acknowledged that she used formal addressing because she did not know her mentor. After more communication, she “accumulated enough information” in CMC and felt a little bit closer to her mentor. She could use a casual way of addressing.

However, the relationship development in CMC was limited by Chinese participants’ mentor-mentee talk. In one Chinese participant’s words, the relationship was “on the surface,” activated by the need to be as casual and comfortable as possible in communication. With mentor-mentee talk, the relationship could never be developed to “friendship” because of the unequal status. One Chinese participant explained about the limited relationship development: “Do we feel closer to each other after dozens of e-mails? Yes, but I still don’t feel he is my friend. I am asking too much and giving too little. It is not equal. We can’t be friends.”
Adoption of Others’ Talk

In intercultural CMC, Chinese participants adopted their U.S. mentors’ talk when they learned the unique online discourse, native English expressions, and U.S. culture. The adoption was encouraged by Chinese participants’ need to learn native language, and Chinese culture underlining solidarity and integration with others. Text-based CMC provided the context that made the adoption possible.

Chinese participants were often confused by the unique English discourse in CMC. After learning the discourse, Chinese participants adopted them in return. In one Chinese participant’s e-mail, she asked about the meaning of “ill talk to ya soon.” In the response, the U.S. mentor explained: “it is … kind of like a simple way to say goodbye online … keep in touch … ill speak with you in the near future … things like that … and ‘ya’ just is slang for ‘you’. ‘ill’ is a abbreviation for ‘I’ll.’ In online talk, you use lower case letters as much as possible.” Interestingly enough, in the next e-mail, this Chinese participant concluded the e-mail with “ill take [talk] to ya soon.” All participants mentioned that they learned to use “live English” through CMC. One participant said that he learned to conclude e-mails with “If you have questions, feel free to ask me.” Whenever he answered his mentor’s questions, he always concluded with this sentence.

In the communication, Chinese participants also learned to adapt to the U.S. culture. When Chinese participants asked about the weather, they had to use degrees. There was confusion between metric and English system. Chinese participants used Celsius and the U.S. mentors used Fahrenheit. One U.S. mentor asked for the conversion to Fahrenheit when a Chinese participant used Celsius: “What is –2 Celsius? Can you tell me what’s in Fahrenheit? Is it cold like –2 Fahrenheit?” The Chinese participant answered back: “–2 Celsius is like 30 degrees Fahrenheit. It is cold. But is not like deadly cold. I think 0 Celsius is about 32 Fahrenheit and 10 Celsius is about 50 Fahrenheit. I will use Fahrenheit from now on. It is easy for me to do the conversion. I got to be ready for the system anyway.”

In interviews, Chinese participants explained their adoption of some English expressions as helping them to learn native English. Such adoption happens in interacting between people of a second language and native speakers. The adoption can be explained by behavioral adjustments people make during communication. It matches communication accommodation that results in favorable social attractiveness, communication effectiveness, and cooperativeness (Giles, Coupland, & Coupland, 1991). On the other hand, Chinese participants also used Chinese collectivism culture to explain their adoption of the U.S. mentors’ talk. As an essential theme in the literature of Chinese use of CMC, Chinese culture has been used widely to explain Chinese communication behaviors (Chen, 1998; Chung, 1992; Giese, 2003; Weber, 2002). Chinese collectivism culture emphasizes solidarity and integration with others, and prioritizes the needs of the group over the needs of the self. Being collectivistic, Chinese people emphasize in-group belonging, personal interdependence, and social harmony with others (Cheng, 1987; Lu, 1997). In one participant’s words, their adoption of the U.S. mentors’ talk led to similarity instead of dissimilarity that was “good for both sides.”

In addition, CMC characteristics allowed Chinese participants to adopt others’ talk. With text-based communication, CMC did not have “intimidating factors,” such as body movements, vocal tones, and facial expressions. With only text in communication, Chinese participants would be free to ask questions about language, learn U.S. culture, and improve their English. One participant said: “In a face-to-face interaction, I would be embarrassed to ask my mentor to explain the meaning of a English word. But in e-mail, I don’t have to care about how my mentor thinks.” In addition, text-based CMC provided the context for Chinese participants to read, reread, and reflect
on the U.S. mentors’ talk before their adoption. Some Chinese participants admitted that in an oral communication, their mentors’ talk would slip away and they would have nothing to learn in the adoption.

**Icebreaker Talk**

Chinese participants were always careful about what they should talk about in CMC with the U.S. mentors. They tended to stay with “general topics” that were safe for both sides. Some participants termed this talk as “icebreaker talk.” Chinese participants defined “icebreaker talk” as “talk about the surface,” “talk as contact,” and “talk without specifics.” In an icebreaker talk, Chinese participants picked up general topics and never went into specifics. For example, one Chinese participant asked about textbooks: “Would you like to tell me where I buy textbooks? Are they expensive?” In the interview, the participant explained that she did not want to ask for specifics, such as textbooks in her major, exact prices for textbooks, required textbooks, and optional textbooks. She did not expect specifics over CMC. She said that the question functioned more as a way to initiate and maintain the communication with her mentor.

Chinese participants attributed their icebreaker talk to different cultural understandings of CMC. In the U.S. culture, e-mail as well as other formats of CMC become one part of routine communication practice. However, in China, e-mail and other formats of CMC are not popular as ways of communication. Guo (2005), along with other colleagues in the Chinese Academy of Social Sciences, conducted a survey tracking Internet use in five major Chinese cities. According to the survey, over two-thirds of people never or seldom use e-mail, instant messaging, online discussion, and chat groups as ways of communication. Close to 40% of the survey participants check e-mail once or less than once a week. Most survey participants (about two-thirds) use CMC on the Internet for news, especially entertainment-related news and online games. Not surprisingly, Chinese participants in this study mentioned that they would prefer face-to-face interaction to CMC as a way of communication. One Chinese participant offered an example about a typical communication event in China: “If you have questions or problems, you need meet people. Sit down with a couple of tea. Take it easy. Negotiate back and forth.” All participants thought that different formats of CMC functioned as a good way to initiate a contact, but face-to-face interaction was the best for further communication.

Chinese participants also talked about the effect of CMC characteristics on their icebreaker talk. All participants agreed that e-mail or instant messaging functioned well at the beginning between them and their mentors. In one participant’s words, e-mail served as an icebreaker. Without context information, it was easy for Chinese participants to begin the communication with their U.S. mentors. After the beginning, e-mail turned out to be a constraint between them and their mentors. Most participants wished that they could have face-to-face interaction after a few weeks of e-mail exchange.

**FUTURE TRENDS**

**Summary of the Study**

In this study, there were four language patterns in CMC between Chinese participants and their U.S. mentors: others-oriented talk, mentor-mentee talk with limited relationship development, adoption of others’ talk, and icebreaker talk. The four language patterns were affected by both cultural and CMC effects. First, in others-oriented talk, Chinese participants showed extreme politeness, considered others as the priority, and talked about others in communication. Chinese culture about the significance of personal relationships formulated others-oriented talk. CMC constrained
others-oriented talk with its directness and written formality. Second, in mentor-mentee talk, Chinese participants talked according to their role as mentees. Chinese participants respected their mentors because of the Chinese meaning of “mentor” as “teacher” or “consultant.” CMC only allowed a limited relationship development between Chinese participants and their U.S. mentors. Third, in the adoption of others’ talk, Chinese participants learned the unique online discourse, native English expressions, and the U.S. culture. Both Chinese participants’ need to learn native language and Chinese collectivism culture encouraged Chinese participants to accommodate to others’ talk. With its text-based communication, CMC allowed Chinese participants’ learning and adoption of their U.S. mentors’ talk. Fourth, Chinese participants regarded CMC as a good icebreaker in intercultural communication. The lack of context cues made CMC appropriate for icebreaker talk and at the same time constrained further communication.

**Discussion of Implications of Research for Current Situations**

In each of the four language patterns of Chinese participants’ communication, cultural values and effects play a significant role. For others-oriented talk, it is Chinese culture about the significance of personal relationships. For mentor-mentee talk with limited relationship development, it is Chinese understanding of the term “mentor” as “teacher” or “consultant.” For adoption of others’ talk, it is Chinese participants’ need to learn the native language and Chinese collectivism culture that encourage communication accommodation. For icebreaker talk, it is CMC’s unpopularity as a way of communication in China and Chinese preference for face-to-face communication.

Chinese culture that is under the strong Confucian influence becomes essential to explain Chinese use of CMC on the Internet. When Bockover (2003) evaluated the impact of Confucian values, she argued that traditional Chinese values are still strong in Chinese use of CMC. Critical to Confucian values is the concept of a person that is always an essential part of a larger social group, and “personal agency” is always socially defined (Bockover, 2003, p. 164). Thus, in China, the main moral goal has been harmonious interdependence in personal relationships. Bockover argued that the Internet with its autonomy, independence, and free trading of information could be at odds with Chinese cultural values. In the study of the dynamics and political function of the transnational Chinese cultural sphere on the Internet, Yang (2003) argued that cultural repertoire with Chinese values and traditions regulates communications and actions of Chinese participants from mainland China, Hong Kong, Taiwan, and North America. In the study of Chinese policy making and implementation on the Internet, Weber (2002) defined Chinese culture as “diffuseness culture” that emphasizes relationship closeness and harmony in social structure, as opposed to Western individualistic “specificity culture.” These cultural differences lead to interpretation problems of the Internet policies in China. Adding to the literature, the findings of the study offer empirical evidence for the impact of culture on intercultural CMC.

On the other hand, the characteristics of CMC function as either a constraint or a way of encouragement in this study. The supplemental effect of CMC characteristics contradicts the findings about CMC’s determining effect on intercultural CMC (Ma, 1996; Tella & Mononen-Aaltonen, 1998). For example, among Ma’s (1996) widely cited five propositions about intercultural CMC, two propositions are directly derived from CMC’s determining effect. First, East Asians tend to show greater self-disclosure in CMC conversations than in face-to-face conversations. Second, status difference is unnoticeable in intercultural CMC. In the current study, Chinese participants’ icebreaker talk about “safe” and “general” topics is contradictory to the proposition about more
Intercultural Computer-Mediated Communication Between Chinese and U.S. College Students

self-disclosure in CMC. Chinese participants’ talk from the mentor-mentee relationship does not match the proposition about unnoticeable status difference in intercultural CMC. One explanation is the situational use of CMC that varies from culture to culture, context to context, and people to people (Warschauer, 2000). Another explanation may be about Ma’s study sample that consists of Chinese students attending U.S. universities and Taiwan universities. Under the U.S. cultural impact, Chinese students attending U.S. universities may not incorporate much Chinese culture in intercultural CMC. Even Chinese students attending Taiwan universities may not incorporate much Chinese culture in CMC because of strong globalization in Taiwan. In the current study, Chinese participants have never communicated with native speakers in such a scope before. These Chinese participants showed much more cultural effect on their language patterns in CMC.

Chinese culture underlines the patterns of communication that Chinese students have with their U.S. mentors. The patterns can be explained by Philipsen’s (1992) theory of two codes in speaking: a code of honor and a code of dignity. Chinese participants communicate with their U.S. mentors from a code of honor. In a code of honor, “the individual is conceived a role or character, a persona” (Philipsen, 1992, p. 109). In the communication with the U.S. mentors, Chinese participants kept the roles of mentees and perceived their U.S. counterparts in the roles of mentors. In a code of honor, society along with others exists before an individual. In this study, Chinese participants had others-oriented talk and adopted others’ talk for similarity and integration with others. On the virtual stage of intercultural CMC, Chinese participants used the talk in CMC as an icebreaker contact instead of actual communication. In a code of honor, Chinese participants viewed their talk as fulfilling their role as a mentee in the social order with their U.S. mentors.

On the other hand, the U.S. mentors communicate with Chinese participants from a code of dignity. In U.S. culture, a code of dignity is “concerned with the person qua person, as someone who is made up of unique feelings, ideas, and attitudes” (Philipsen, 1992, p. 113). In a code of dignity, communication functions to express individual intent and will. The U.S. mentors engaged in individual talk that featured personal hobbies, personal interests, and personal plans. The U.S. mentors also engaged in friendship talk, treating Chinese participants as equal individuals in intercultural CMC.

Implications of Research for Future Situations

The findings in the study call attention to social constructivism in future studies of communication technologies. In literature about communication technologies, social constructivism advocates that people in a particular culture shape social behavioral patterns regarding technology use (Fischer, 1992; Fulk, 1993; Fulk & Boyd, 1991). Through negotiations and struggles among people, the meanings in people’s use of technologies are internally constructed and redefined in changing cultural contexts (Barley, 1990; Leonard, 2003; Xia, 2006). The application of social constructivism ideas calls for more analysis of technology use for a particular culture in future studies. The findings about the determining cultural effect on Chinese participants’ use of English in CMC with their U.S. mentors offer an example of cultural analysis of technology fit from a social constructivism perspective. In the current study, the four language patterns of Chinese participants’ communication call for the evaluation of a particular culture as a part of social construction process.

Suggestions for Future Applications of the Findings

It is interesting to consider the U.S. students’ responses in discussing this chapter’s future applications of to intercultural CMC. Some U.S.
students felt strange with Chinese participants’ talk about families, friends, and classmates. Some U.S. students felt annoyed by Chinese participants’ paragraphs of apology. Some U.S. students did not understand Chinese participants’ “touch and run” talk. These reactions can be explained by the discrepancy in self-perceptions between the East Asian and U.S. students. Even if the East Asians thought of themselves as direct in CMC, the U.S. students still found them extremely polite, reserved, and indirect (Ma, 1996). How to change this misunderstanding should be the focus of future studies.

The findings in the current study uncover the importance of understanding cultural resources regarding language activities in intercultural CMC. The assumed relation between one language and one culture explains significant meanings in Chinese participants’ communication with the U.S. students. Although Chinese participants use English in CMC, the rules and meanings regarding language come from their native culture. As Hymes (1968) proposed in the assumption about speaking in a cultural group, Chinese participants’ use of English language in CMC consists of a unique system, and their actual language activities should be the focus of study. Knowledge of cultural resources about language activities is the key to mutual understanding. The findings in the current study about Chinese culture in Chinese participants’ others-oriented talk, mentor-mentee talk, adoption of others’ talk, and icebreaker talk can help their U.S partners as well as people from other cultures understand diverse cultural resources in the future intercultural communication over CMC.

**Suggestions for Future Studies**

With the fast development and diffusion of CMC technology as well as cultural globalization, interaction in CMC across different cultures is becoming more and more common. English as a global language is the key for intercultural CMC. As Cassell and Tverksy (2005) reviewed, studies of the particular use of online English in different cultures are sparse. The future research of intercultural CMC should focus more on the particular rules, patterns, and meanings of English language use in different cultures and uncover cultural resources regarding English language activities in CMC.

In the study, there are 12 Chinese participants. Although small samples are common for in-depth qualitative studies, the generalization of the findings to other Chinese students is unknown. The relation between cultural resources and language use in CMC may not be generalized to other groups of people. Further examination with large samples is needed for other groups. Discussing different language patterns in CMC individually helps to describe themes, patterns, and ideas clearly. However, it does not represent the interconnections among them. Apparently, Chinese participants’ others-oriented talk can affect their adoption of others’ talk. More research is needed to unpack these interconnections among different language patterns in Chinese use of English in CMC.

**CONCLUSION**

The study incorporated both cultural effects and CMC characteristics in the investigation of Chinese college students’ use of English in communication with U.S. college students. Ethnography of communication was used as the method to uncover four language patterns in Chinese college students’ use of English: others-oriented talk, mentor-mentee talk with limited relationship development, adoption of others’ talk, and icebreaker talk. Chinese cultural values showed strong effect on Chinese college students’ use of English in CMC. Chinese cultural values included Chinese significance of personal relationships, Chinese collectivism, Chinese understanding of authority, and Chinese use of CMC. The characteristics of CMC showed minimal or supplemental
effect as either a constraint or encouragement in the four language patterns.

REFERENCES


This work was previously published in Linguistic and Cultural Online Communication Issues in the Global Age, edited by K. St.Amant, pp. 63-77, copyright 2007 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 5.9
Global Culture and Computer Mediated Communication

Susan R. Fussell
Carnegie Mellon University, USA

Qiping Zhang
Long Island University, USA

Leslie D. Setlock
Carnegie Mellon University, USA

ABSTRACT

In this chapter, we discuss how culture influences computer mediated communication (CMC). We use an Input-Process-Output (I-P-O) model as a theoretical framework to analyze relationships between culture and CMC. We describe three dimensions of cultural variability—individualism/collectivism, low vs. high context of communication, and task- vs. relationship-orientation—and describe how these dimensions influence people’s reliance on features of CMC. A review of the literature to date suggests that cultural factors do indeed shape how people use CMC. More specifically, auditory and visual cues appear to have more importance for members of collectivistic, high-context, relationship-oriented cultures than they do for members of individualistic, low-context, task-oriented cultures. However, further research is needed to clarify relationships between cultural dimensions and CMC, to understand the role of moderating variables such as gender of participants, task, and group composition, and to provide design guidelines for new tools to support intercultural communication and CMC in developing regions.

INTRODUCTION

One of the great benefits of the CMC tools covered in this volume is that they allow people to converse across space and time. Today, people connect with others from around the world by participating in chatrooms and discussion lists, by joining global game communities and virtual worlds, by authoring and reading blogs with an international scope, and by a variety of other
Global Culture and Computer Mediated Communication

means. In the work domain, firms are increasingly taking advantage of CMC tools to establish global teams with members from a diverse set of nations (e.g., Carmel, 1999; Churchill & Bly, 2000; Olson et al., 1998). Bridging nations via technology does not, however, guarantee that the cultures of the nations involved are similarly bridged (Olson & Olson, 2000). Mismatches in social conventions, work styles, power relationships and conversational norms can lead to misunderstandings that negatively affect the interaction. For example, an individual from a task-oriented culture such as the United States or Canada may focus exclusively on getting things done, overlooking the social niceties expected by his/her conversational partner from a relationship-focused culture such as China, Japan or Latin America. Similarly, an individual from a low-context communication culture, who relies primarily on verbal language to express his or her thoughts, may ignore facial expressions or tones of voice that are intended to be communicative by his/her partner from a high-context culture.

As the examples above suggest, features of different CMC tools (e.g., e-mail, Instant Messaging [IM], teleconferencing, video conferencing) may make it easier or harder for people to bridge their cultural differences by providing or failing to provide auditory, visual and other sources of information. But how can we characterize these differences between CMC tools and use them to decide what tools will work best for a given set of people performing a given set of tasks? For members of Western cultures, years of research have led to a number of well-developed theories that can be used to motivate such decisions (e.g., Clark & Brennan, 1991; Daft & Lengel, 1984; Postmes, Spears, & Lea, 2002; Short, Williams, & Christie, 1976; Daft & Lengel, 1984, Clark & Brennan, 1991; Walther, 1992, 1995; Postmes, Spears & Lea, 2002). Much less is known about CMC in intracultural (culturally homogenous) groups from non-Western cultures or in intercultural (culturally heterogeneous) groups, making it harder to predict the suitability of different media for different purposes or to design new CMC technologies to facilitate intercultural collaboration.

There are reasons to believe that what we know from research using Western participants will not generalize straightforwardly to other cultures. Cultures vary along a number of dimensions that may impact group processes and outcomes, such as individualism vs. collectivism (the extent to which people prioritize their personal benefit vs. that of the larger group; e.g., Hofstede, 1983; Triandis, 1995), low- vs. high context of communication, (how much contextual information is required for communication; Hall, 1976), and task vs. relationship orientation (whether people focus on getting work done or on establishing rapport with their partners; e.g., Triandis, 1995). These and other cultural dimensions may interact with features of media to create different effects on group processes and outcomes than have been found in studies using Western participants. For example, if expressions and intonation are important parts of communication in high context cultures, standard IM, which doesn’t support seeing or hearing one’s partner, should be less suitable for communication than it is in low context cultures. Consistent with this argument, recent research does in fact suggest that people’s cultural background affects CMC (e.g., Anderson & Hiltz, 2001, Kayan, Fussell, & Setlock, 2006, Reinig & Mejias, 2003, 2004; Setlock, Fussel, & Neuwirth, 2004; Setlock, Quinones, & Fussell, 2007; Zhang, Olson, & Olson, 2004; Zhang, Sun, Chintakovid, Ge, Shi, & Zhang, 2006). Although these studies vary widely in terms of what technologies and cultures are studied and what research methods are used, we can use the results to begin piecing together a theoretical framework for understanding relationships between culture and CMC.

The objectives of the chapter are to provide an overview of research in the area of culture and CMC to date. We first present a conceptual framework for understanding how culture and
CMC shape communication processes and task outcomes. Then, we review research on each component of this framework, highlighting findings that with broad implications for research in CMC. We conclude with some directions we think are especially important for future research.

BACKGROUND

The theoretical framework we use to examine relationships between culture and CMC is an Input-Process-Output (I-P-O) model (e.g., Hackman, 1987), shown in Figure 1. Here, culture and media are inputs that people bring to collaboration. These inputs, both alone and in interaction, influence communication processes and, in turn, subjective and objective outcomes. There are also a number of moderating variables that may influence relationships between inputs and processes and between processes and outcomes. Although the I-P-O framework is a simplification, it can help us conceptualize how culture and CMC interact by explicating relationships between input, process, outputs and moderating variables.

Dimensions of Cultural Variability

The first input to our model is the cultural values of the communicators. For the purposes of this chapter, we define culture as a set of norms, roles and values emphasized by a culture and adopted, to greater or lesser degrees, by members of that culture through such processes as imitation and teaching. Although researchers agree that cultures differ in many ways, they have debated how culture should be defined and the specific dimensions upon which cultures vary (e.g., Hofstede, 1983; Oyserman, Coon, & Kemmelmeier, 2002; Schwartz, 1992; Triandis, 1995). We focus our discussion on three cultural dimensions—individualism vs. collectivism, high vs. low context of communication, and task vs. relationship focus—that are especially important to CMC.

Individualism vs. collectivism. Many culture theories distinguish between individualistic and collectivistic cultures. In individualistic cultures, people tend to identify themselves first and foremost as individuals, with the primary goal of personal (or perhaps familial) gain. In collectivistic cultures, people tend to identify themselves as a member of a larger collective and focus on the betterment of that collective (e.g., Hofstede, 2001; Triandis, 1995). Whether someone is from an individualistic or collectivistic cultural background has far-reaching effects, influencing his or her cognitive processes (Nisbett, 2003) and his/her self-concept as either independent of or interdependent with other individuals (Markus...
Global Culture and Computer Mediated Communication

Table 1. Some affordances of communication media and their typical presence (Y), partial presence (P) or absence (N) in face-to-face communication (FtF), video conferencing, telephone, and instant messaging (IM) (Adapted from Clark & Brennan, 1991)

<table>
<thead>
<tr>
<th>Affordance</th>
<th>Definition</th>
<th>FtF</th>
<th>Video</th>
<th>Phone</th>
<th>IM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audibility</td>
<td>Participants hear other people and sounds in the environment.</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Visibility</td>
<td>Participants see other people and objects in the environment.</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Copresence</td>
<td>Participants are mutually aware that they share a physical environment</td>
<td>Y</td>
<td>P</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Cotemporality</td>
<td>Participants are present at the same time.</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

& Kitayama, 1991). In addition, Hofstede’s (2001) analyses of survey responses from IBM employees around the world show that cultural tendencies toward individualism vs. collectivism are associated with many aspects of daily life, including business practices, child-raising, and educational techniques.

High- vs. low-context of communication. Another way that cultures have been argued to differ is in the strategies they use to communicate. Hall (1976) distinguished between low context communication, which is verbally explicit, to the point, with relatively little attempt to mask one’s feeling and high context communication, which is indirect, often ambiguous, and sensitive to the specific situational context (e.g., the relationship between speaker and addressee, nuances of facial expressions or tone of voice). For example, low context communicators are likely to disagree outright with their conversational partners, whereas high context communicators may use silence or indirect speech to indicate disagreement. Although people in all cultures use both communication styles, research suggests that low-context communication is preferred in individualistic societies such as the U.S. and Canada and high context communication is preferred in collectivistic societies, particularly Asian cultures (Gudykunst & Ting-Toomey, 1988; Gudykunst et al., 1996).

Task vs. relationship focus. A third way that cultures have been argued to differ is in the orientation they take toward the task at hand (e.g., Triandis, 1995). Task-oriented cultures focus on getting work done, whereas relationship-oriented cultures focus on establishing rapport with one’s partners. As we noted earlier, these differences in orientation may lead to misunderstandings and misattributions in collaborative work. When individuals from task-oriented cultures fail to attend to relationship factors, they may be perceived as rude by their teammates from relationship-oriented cultures; when individuals from relationship-oriented cultures fail to focus on the task, they may be perceived as undedicated by their teammates from task-oriented cultures. The task vs. relationship focus is only quasi-independent of the other dimensions. Individualistic cultures such as the U.S., Canada, and Australia tend to use low-context communication styles and have a task orientation. Collectivistic cultures such as Japan and China tend to use high-context communication styles and to have a relationship orientation.

Although dimensions like individualism-collectivism are often applied at the national level (e.g., Hofstede, 1983, 2001; Triandis, 1989), we almost must consider how these national values are reflected in individuals’ personal values (e.g., Schwartz, 1992) and their construals of themselves as interdependent vs. independent (e.g., Markus & Kitayama, 1991; Singelis & Brown 1995). Both individual values and individuals’ self-concepts are influenced by national cultural but not entirely determined by it. Gudykunst et al. (1996) and Oetzel (1998), among others, have
shown that national values, individual values, and self-concepts all influence the ways people communicate.

**Affordances of Media**

The second input to our model is the medium used for communication. More specifically, we are interested in how features of CMC tools might interact with culture to shape communication processes and task outcomes. We draw on Clark and Brennan’s (1991) influential theory of *media affordances* as a framework for understanding differences among media (see Table 1). For example, telephone calls and video conferencing provide audibility, and thus afford the use of vocal speech production, whereas IM does not. In this framework, features of media alter the costs or difficulty of producing messages, receiving and understanding messages, changing speakers, and repairing misunderstandings.

**APPLYING THE FRAMEWORK TO COMMUNICATION AND TASK OUTCOMES**

Starting from Clark and Brennan’s model, our goal is to understand how cultural dimensions may alter the perceived importance of audibility, visibility, copresence and other media affordances. For example, Gudykunst and Kim (1997) suggest that nonverbal cues such as gestures and facial expressions may be more important for communication in high-context cultures. Thus, we might anticipate that visibility will be more important for members of high-context cultures than for members of low-context cultures. In this section we examine how culture and media influence conversational processes and team outcomes.

**Communication Processes**

Culture may, alone or in interaction with features of media, influence group processes, particularly processes of communication. We see two aspects of communication as especially relevant for CMC: conversational grounding and relational communication. In this section we review prior work on the ways in which culture and media affect these two communication processes.

**Conversational Grounding**

*Conversational grounding* refers to the way people interact to ensure that they understand one another’s messages (Clark & Brennan, 1991; Clark & Wilkes-Gibbs, 1986). Speakers and listeners ask each other questions, provide clarifications, rephrase one another’s messages, and so forth. The purpose of these activities is to ensure that messages are understood as intended. Grounding is easier, and conversation more efficient, when collaborators share common ground, mutual knowledge, beliefs, and so forth (Clark & Marshall, 1981). This common ground can arise from comembership in social groups (e.g., Fussell & Krauss, 1992; Isaacs & Clark, 1987), through the process of exchanging messages (*linguistic copresence*), or by sharing a physical setting (*physical copresence*).

![Figure 2. Mean speaking turns per task by culture group and medium (AA = American only, AC = Mixed American Chinese, CC = Chinese only)](attachment:image)
Affordances of media influence the strategies people use to ground their utterances (Clark & Brennan, 1991; Fussell et al., 2004; Gergle, Kraut & Fussell, 2004; Kirk & Stanton-Fraser, 2006; Kraut, Fussell, & Siegel, 2003). For example, face-to-face settings afford visibility and physical copresence, so speakers can use gestures and deictic expressions such as *this one* to refer efficiently to objects and people in the environment (e.g., Bekker, Olson, & Olson, 1995; Clark & Krych, 2004). On the telephone or over IM, media that lack visibility and physical copresence, speakers must use lengthier verbal descriptions of the same objects (e.g., Doherty-Sneddon et al., 1997; Kraut et al., 2003).

Of particular interest for building theories of culture and CMC is the idea that cultures vary in their strategies for grounding conversations (e.g., Li, 1999a, 1999b). Hall (1976) proposed that audibility and visibility were more important for grounding in high-context cultures than in low-context cultures, because awareness of how others are reacting to one’s messages is an important aspect of high-context communication. This notion is supported indirectly by Veinott, Olson, Olson, and Fu (1999), who found that nonnative English speakers, many of whom were Asian, benefited from video over audio conferencing, whereas native English speakers did not. Veinott and colleagues suggest that the richer cues to mutual understanding provided by visibility (e.g., quizzical looks, raised eyebrows) were especially valuable for nonnative speakers, but it is equally plausible that video was important because these participants were from high-context cultures.

Our own prior studies suggest a less clear-cut relationship between the availability of auditory and visual cues and conversational grounding in low- vs. high-context cultures. In our first study (Setlock et al., 2004), we compared American, Chinese, and mixed American-Chinese dyads performing a negotiation task face-to-face or via IM. We hypothesized that the lack of social/contextual cues in IM would make it poorly suited for communication among members of high-context cultures, whereas it should have minimal or no effect on communication among members of low-context cultures. Consistent with this hypothesis, we found no difference between media in terms of how much grounding American pairs required to complete the task, but a large culture by medium interaction such that Chinese pairs spoke much more face-to-face (Figure 2).

Face-to-face interaction has many affordances not present in IM, including audibility and visibility. To determine which of these two affordances was more important, we (Setlock et al., 2007) compared the same cultural groups interacting via audio or video conferencing. In this case, however, we found no main effect of culture, nor a culture by medium interaction. Thus, audibility seems to be important for Chinese dyads, but adding visual cues via video conferencing did not provide additional benefit. These results conflict with those of Veinott et al. (1999), reported above, and suggest the need for a more detailed examination of factors that differed across the two studies (e.g., tasks, specific cultural backgrounds of participants).

Relational Communication

Relational aspects of communication are concerned not with *what* information is conveyed but with *how* that information is conveyed and what this indicates about the relationship between speaker and addressee(s). For example, nonverbal cues such as eye gaze (e.g., Argyle & Cook, 1976), facial expressions (e.g., Ekman, 1982), and posture (e.g., Mehrabian, 1967) can be used to express intimacy, trust, and attraction. In addition, the verbal content of messages can be crafted in different ways to establish, maintain, or build closeness with (or to maintain or increase distance from) a partner. Speakers’ decisions about forms of address, informal vs. formal language, use of swear words, and so forth, can all have implications for their relationships with their addressees. In Table
Global Culture and Computer Mediated Communication

Table 2. Basic categories of relational communication

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonverbal behavior</td>
<td>Eye gaze, posture, facial expression</td>
</tr>
<tr>
<td>Paralinguistic behavior</td>
<td>Intonation patterns, speech rate, loudness</td>
</tr>
<tr>
<td>Verbal behavior</td>
<td>Form of address (e.g., John, Mr. Jones) pronouns (e.g., I, we), hedges (e.g., sort of), intensifiers (e.g., very very), indirect requests (e.g., would you mind …), swear words</td>
</tr>
</tbody>
</table>

2, we list some of the ways that nonverbal, paralinguistic and verbal cues can add socio-affective meaning to people’s messages.

Which features a medium possesses affects people’s relational communication (e.g., Herring, 1994; Kiesler, Siegel, & McGuire, 1988; see Whittaker, 2003 for a review). For example, some studies have found that when audibility and visibility are absent, as in IM, people are more likely to “flame” their partners (e.g., Kiesler et al., 1988). Kiesler and colleagues attributed this finding to the lack of social context cues in text communication: it is easier to be insulting when you can’t see others’ reactions to the insult. Other investigators have argued that it might instead take more effort to produce the extra words needed to be polite (e.g., sort of, would you mind?) in typed discourse (e.g., Brennan & Ohaeri, 1999), resulting in speakers eliminating those seemingly nonessential words in the interest of efficiency. Other studies have found more relational communication in face-to-face settings than over the phone (Rutter, 1987; Stephenson, Aylng, & Rutter, 1976) or in text communication (e.g., Hiltz & Turoff, 1978). Less research has compared relational communication in audio vs. video conferencing, although many media theories (e.g., Daft & Lengel, 1984; Short et al., 1976) suggest that video will better support relational communication.

The cultural theories discussed earlier suggest that high-context, relationship-oriented cultures place more emphasis on relational communication than do low-context, task-oriented cultures (Ting-Toomey et al., 1991). One area in which this hypothesis has been tested is that of conversational indirectness. People can ask others to do things in a variety of ways, including direct commands (e.g., take out the trash), indirect requests (e.g., would you mind terribly taking out the trash?), and off-record statements (e.g., it’s terribly messy in here). In general, the less direct the request, the more polite it is (Brown & Levinson, 1987). Cross-cultural studies of conversational indirectness have shown that speakers from high-context cultures such as China and Korea use more indirectness than those from low-context cultures such as the United States (Ambady, Koo, Less, & Rosenthal, 1996; Holtgraves, 1997). Chinese speakers are also more likely than American speakers to use language that promotes relationship-building, such as “we” pronouns and social language, both face-to-face and via IM (Setlock et al., 2004). Cross-cultural research on negotiation, in which people work to resolve disagreements, has shown that relational strategies (e.g., compromising) are favored by high-context negotiators whereas informational strategies (e.g., dominating the conversation) are favored by low-context cultures (e.g., Adair & Brett, 2005; Adair, Okumura, & Brett, 2001; Ting-Toomey et al., 1991). Such differences have been attributed to cultural variation in concern for one’s own face vs. the other person’s face (Ting-Toomey, 1988).

Based on these considerations, we would expect a person’s cultural background to interact with the features of a communication medium to influence the production of relational communication. For low-context communicators, a medium such as IM that requires verbal substitutes for relational features such as facial expressions, tone of voice, and gestures may be perfectly suitable. For high-context communicators, who rely more heavily on nonverbal communication, media that are lacking auditory and visual cues may disrupt important aspects of social interaction.
Outcome Measures

In our I-P-O model in Figure 1, inputs (culture and features of technology) impact communication processes such as conversational grounding and relational communication. These communication processes in turn impact a range of outcome measures. Here, we touch briefly on several outcomes that are especially important in intercultural teamwork: perception and subjective evaluations of one’s partners, persuasion, trust, and objective task performance.

Perception of partners. Features of media affect collaborators’ impressions of each other’s personality, intelligence, physical attractiveness, and so forth. For example, people like others more when they can see them as well as hear them (Short et al., 1976; Williams, 1977). Hancock and Dunham (2001) suggest that the lack of social cues in text CMC creates ambiguity. Is a partner being rude, or just a bad typist? Does a lack of response mean that a partner is being rudely dismissive or that his/her computer crashed? Cramton (2001) has shown that when people know less about a remote collaborator’s immediate experiences, they are more likely to attribute problems to internal, dispositional factors (e.g., a partner’s rudeness) than to external causes such as computer problems (Cramton, 2001). Such effects may be especially strong in intercultural interaction, in which people share less initial background knowledge. In our own studies (Setlock et al., 2004, 2007), we found that members of intercultural dyads rated each other more negatively than members of homogenous American or Chinese dyads. However, the tendency to attribute behavior to dispositional factors is in part culturally-specific: East Asians are more likely to consider situational explanations for behaviors than Westerners (e.g., Choi, Nisbett, & Norenzayan, 1999; Morris & Peng, 1994). As a result, we would predict that media lacking visual and auditory cues would have a greater impact on partner perceptions in high-context than low-context cultures. A study currently underway in our laboratory is testing this hypothesis.

Persuasion. Persuasion refers to the extent to which a person can convince others that his or her viewpoint is correct. Early studies indicated that persuasion varied as a function of medium (e.g., Chaiken & Eagly, 1976; Guadagno & Cialdini, 2002; Heim, Asting, & Schliemann, 2002; Morley & Stephenson, 1977), but many of these studies used artificial role-playing paradigms in which conversational grounding and relational communication could not be measured. Other studies have compared how much consensus groups reach using text-based group decision support systems (GDSS) and found less consensus after GDSS than after face-to-face negotiations (Reinig & Mejias, 2003; Watson, Ho, & Raman, 1994). There is little agreement as to whether cultural differences in negotiation styles influence persuasion, either alone or in interaction with features of technology. Some studies (e.g., Reinig & Mejias, 2003; Watson et al., 1994) have found that both culture and medium affect persuasion but no interaction between the two. Others (e.g., Setlock et al., 2004) have found effects of culture but no effects of medium. Adair et al. (2001) suggest that persuasion may be reduced when there are mismatches in negotiators’ styles, although they did not investigate interactions with medium. That there is still much to learn about culture and CMC is suggested by a study by Anderson and Hiltz (2001), in which they found—in contrast to the other studies we’ve reviewed—that face-to-face culturally heterogeneous groups showed the most consensus and asynchronous culturally-homogeneous groups showed the least consensus after group discussion.

Trust. Trust refers to a person’s confidence in the goodwill of others and his/her expectation that others will reciprocate if he or she cooperates (e.g., Ring & Van de Ven, 1994). McAllister (1995) differentiates two broad foundations for trust: cognitive and affective. Cognitive trust is built on respect for others’ intelligence, competence and reliability, whereas affective trust is built on emotional bonding and relationship-building. We
would expect affective trust to be weighted more heavily in high-context, relationship-oriented cultures and cognitive trust to be weighted more heavily in low-context, task-oriented cultures.

Establishing trust is more difficult over CMC than face-to-face, and more difficult with leaner media like IM than with richer media like audio or video conferencing (Bos et al., 2002). To some extent, however, the problems establishing trust via CMC can be reduced when people start with an initial face-to-face interaction before working at a distance (Rocco, 1998; Jensen, Farnham, Drucker, & Kollock, 2000). For high-context cultures, characteristics of the communication medium may be especially important for establishing affective trust. Zhang, Olson, and Olson (2004) found that Chinese pairs showed higher affective trust when negotiating by video conferencing than by audio conferencing, whereas American pairs showed no differences on either affective or cognitive trust between media.

Objective performance. Lastly, the success with which people communicate is predicted to influence how well they perform collaborative tasks. This performance can be measured in a variety of ways, such as task completion times, numbers of errors, or quality ratings. Because the time it takes to do a communication-based task is correlated with how much talking is required, media that support efficient communication through spoken language and gesture typically lead to faster performance times (e.g., Gergle, Kraut, & Fussell, 2004; Kraut, Fussell, & Siegel, 2003). When performance is measured in other ways, however, the effects of media are less clear. For example, Doherty-Sneddon et al. (1997) found no differences in the accuracy of map routes described in video vs. audio conferencing; Jackson, Anderson, McEwan, and Mullin (2000) found no differences in the quality of poster designs created with high vs. low video frame rates; and Straus and McGrath (1994) found no differences in performance on idea generation, intellective, or judgment tasks between text-based and face-to-face interaction. It has been argued that tasks involving negotiation and persuasion will suffer more from a lack of visual and auditory cues than tasks requiring less interpersonal finesse (e.g., Daft et al., 1987; Short et al., 1976), although support for this proposition has been mixed (see Whittaker & O’Conaill, 1997, for a review). A few studies have looked at the influence of culture on performance quality. Li (1999a, 1999b) found no differences in accuracy of information transmission in Canadian and Chinese dyads, but significantly poorer transmission in mixed Canadian-Chinese dyads. Adair and colleagues (2001) likewise found that intercultural teams performed more poorly on negotiation tasks. Neither study examined interactions between culture and CMC, however. As with the other outcome measures we have discussed, more research is clearly needed in order to understand how culture and CMC influence task performance.

FUTURE TRENDS

The research reviewed above provides some initial insights into how culture and CMC might interact to shape group processes and performance, but as is clear from our review, there are many gaps in the literature. In this section, we briefly describe four areas for future work that we believe are crucial for moving research in this area forward: theory development, research on the effects of moderating variables, development of tools for intercultural collaboration, and designing CMC for developing regions.

Theory Development

As we have noted, prior work presents an incomplete picture of how culture and technology interact to affect teamwork. Some studies show interactions between culture and medium that are consistent with the cultural dimensions of individualism vs. collectivism, high vs. low
context of communication, and task vs. relationship focus; others show main effects of culture or medium but no interaction between them. In part, differences among studies may be due to the wide variety of tasks, cultures, media, and measures they have used. An important goal for future research is the development of a conceptual framework that explicates the existing pattern of findings. In addition, the cultures examined in CMC research must be broadened. Research to date has focused predominantly on only a handful of Asian cultures. The extent to which these results will generalize to others such as Thailand, the Philippines, Vietnam, and so forth, remains to be seen. In addition, we must move beyond the current emphasis on Asian/Western dichotomy to understand how culture interacts with CMC in other areas of the world.

Effects of Moderating Variables

As shown in the I-P-O framework in Figure 1, there is a range of factors that may moderate how culture and technology interact to influence communication and outcomes. For example, how do individual characteristics such as gender and age influence the relationship between culture and media? Are cultural differences more pronounced for certain types of tasks, or in certain contexts (e.g., home, education, or work)? Are they enhanced or diminished when teams interact over extended periods of time? Prior studies have varied widely in terms of participants’ individual characteristics, tasks and time periods of interaction, making it impossible to identify the impact of any single factor. If we are to develop a theoretical framework for understanding how culture and CMC interact, we must isolate the impacts of each of these factors.

Tools for Intercultural Collaboration

Several studies of culture and CMC have found especially poor communication and outcomes for intercultural teams (e.g., Adair et al., 2001; Setlock et al., 2004). These effects may stem in part from a lack of fit between the needs of intercultural teams and the tools available to support them. Thus, an important focus for future research is the development of new tools that can enhance intercultural teamwork. For example, a new CMC tool might automatically modify messages to be more appropriate for the recipient’s cultural background, similar to real-time translation software (e.g., Yamashita & Ishida, 2006). Alternatively, CMC tools might seek to educate users about one another’s cultural norms, for example, by informing the sender of a message as to why it might be inappropriate given the recipient’s culture.

CMC for Developing Regions

A related domain for future research is the investigation of needs for CMC tools in developing regions, in which information technology infrastructure may be weak and in which local customs may be quite different from those in more developed nations (Brewer et al., 2006). An interesting example of work in this area is the Combadge project (Fankel & Bronberg, 2005). Combadge is a voice communication device that addresses a lack of continuous telephone connectivity and low literacy rates in rural India by storing messages in a queue on the Combadge device until the person enters the range of a transmitter tower, at which point the stored messages are sent on to their intended recipients. The design and evaluation of CMC tools for developing regions is likely to be one of the most exciting yet challenging directions for future research.

CONCLUSION

In this chapter we have presented an Input-Process-Output framework for understanding how culture influences CMC. We described three dimensions of cultural variability—individualism
Global Culture and Computer Mediated Communication

vs. collectivism, high vs. low context of communication, and task vs. relationships-orientation—and discussed how these dimensions may interact with features of CMC tools to influence group communication and team outcomes. A review of the literature to date suggests that cultural factors do indeed shape how people use technology to communicate. More specifically, people from cultures that emphasize nonverbal and contextual aspects of communication are more affected by the visual and auditory affordances of CMC tools than are people from cultures that emphasize the verbal aspects of communication. Future research is needed to flesh out the I-P-O model, to understand how moderating variables influence culture and CMC, and to provide design guidelines for new tools to support intercultural communication and CMC in developing regions.

REFERENCES


Rocco, E. (1998). Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact. In Proceedings of Conference on


Watson, R. T., Ho, T. H., & Raman, K. S. (1994). Culture: A fourth dimension of group support
Global Culture and Computer Mediated Communication


**KEY TERMS**

**Conversational Grounding:** The activities by which speakers and listeners work together to ensure that messages are understood as intended.

**Individualism-Collectivism:** The extent to which members of a culture identify themselves as individuals and focus on their personal gain vs. identify themselves as members of a collective and focus on group gains.

**Input-Process-Output Models:** Models of teamwork that examine relationships between variables people bring with them to an interaction (inputs), the interaction among people (process), and the subjective and objective outcomes of this interaction (output).

**Low- vs. High-Context of Communication:** The extent to which communication in a culture is verbally explicit vs. ambiguous and sensitive to the context in which it occurs.

**Media Affordances:** The properties of CMC tools that permit people to use different communication strategies such as speech vs. typing.

**Relational Communication:** Verbal and nonverbal aspects of how messages are conveyed that provide information about the relationship between speaker and addressee(s).

**Task vs. Relationship Focus:** The extent to which members of a culture orient themselves toward getting work done vs. building good relationships with their colleagues.

This work was previously published in *Handbook of Research on Computer Mediated Communication*, edited by S. Kelsey; K. St.Amant pp. 901-916, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 5.10
Linguistics of Computer-Mediated Communication: Approaching the Metaphor

Rosanna Tarsiero
Gionnethics, Italy

ABSTRACT
This chapter introduces the embodied metaphor as a means of studying the multifaceted relationship between computer-mediated communication (CMC) and culture. It argues that the embodied metaphor offers a more reliable framework, as opposed to both deterministic and dispositional perspectives, in that it provides a way to understand, explain, and frame the user’s adaptive response to CMC. Furthermore, it also argues that culture has the potential to shape interactions by influencing the way metaphors are enacted. The author hopes that the aforementioned approach will provide a new understanding of the interplay between CMC and culture, informing future study design and bridging the existing gap between qualitative and quantitative research through action science.

INTRODUCTION
Since its inception, the study of language has been one of the most fascinating disciplines, combining neurophysiology, anthropology, social psychology, and sociology. In recent years, linguistics has been challenged by the advent of the “new media” whose sensory channel permutations are so high and whose sociological and psychological characteristics are so complicated and intertwined that they stress the very concept of “media discourse.”

With the diffusion of the Internet for civil usage, people who had become familiar with using computers for computational, educational, and leisurely purposes also started engaging in human-to-human interactions through machines. Its remarkable simplicity, customization possibility, and flexibility quickly led e-mail to be the killer
application of the Internet all over the world (ITU, 2001; Pew Internet and American Life Project, 2003). More recently with the birth of the Web 2.0 paradigm, the blending of old media into “new” hybrid spin-offs accelerated remarkably. Even text-written computer-mediated communication now has its sub-genres.

The spreading of the Internet and computer-mediated communication (CMC) also generated interest in computer science and social psychology researchers, who tried to make sense of it. However, in the attempt to provide an explanation, several technical, psychological, and social paradigms blossomed, yet none of them explicitly addressed the increasingly higher proportion of cross-cultural exchanges that had started happening after the Internet spread beyond where it was invented. Despite the diffusion of the Internet and CMC, the raising distributed location of their users, the increasing use of e-mail as a collaborative tool in distributed teams, and the amount of studies and information already collected, laymen and scholars alike neglect the role and weight culture has in shaping e-mail exchanges, and assume unrealistic stands that range from extreme determinism to extreme subjectivism.

The purpose of this chapter is to unfold the relationship between CMC and culture through language by exploring the metaphors behind CMC and how they are enacted in different cultural settings. In doing so, the author openly acknowledges and espouses Lakoff and Johnson’s (2003) view of metaphors as cognitive schemes, and relativism as the interplay between subjectivism and objectivism.

In the “Background” section, the literature on CMC, on CMC linguistics, and on cross-cultural studies will spot some reasons of methodological concern. In the “Main Thrust of the Chapter” section, I will explore the metaphors behind CMC and look at them through the lenses of the literature in order to build a model of how culture affects spontaneously occurring computer-mediated exchanges. Finally, in the “Future Trends” section, I will discuss future scenarios and the limitation of the model, suggesting areas for possible refinement of the model and focus of future research.

BACKGROUND

Computer-mediated communication (CMC) and its linguistics were not specifically investigated with regard to how culture impacts them and their interplay. Therefore, this section is composed of three literature reviews: (a) works on CMC, (b) studies on CMC linguistics, and (c) cross-cultural research.

Computer-Mediated Communication

CMC has been extensively studied from psychological, sociological, and technological perspectives. Quantitative approaches outnumbered qualitative ones (Liu, 2002) and suffered with several methodological issues (Culnan & Markus, 1987; Garton & Wellman, 1995; Lamerichs & Molder, 2003; Walther, 1992, 2002). They focused on short-term interactions in artificially constituted groups whose members participated in the research project due to extrinsic rewards (such as higher course grade, extra-credit, money, etc.) or top-down imposition (such as work-related mandated task, workplace or coursework assignment completion, etc.), therefore the research design often became a confounding factor (Walther, 2002). In many of the aforementioned studies, participants were also asked to evaluate the effects of CMC within close categories (Lamerichs & Molder, 2003), pigeon-holing data range, participant experiences, and interpretative schemes in the name of a strong bias in favor of “objectivity.” Often, computer literacy went unaddressed/unassessed, even though it is known to adversely affect interaction outcome (Walther, 2002). Finally, in most quantitative works neither the culture of researchers nor participants was assessed and/or
planned for in the design of the study. Despite all of these biases, the findings of these studies were surprisingly generalized to represent something “intrinsic to the brain” and CMC to be “clearly” inferior and devoid of advantages. Characteristically, quantitative research negative findings are attributed to negative characteristics of either the communication medium (technological determinism, epitomized by Daft & Lengel, 1984; Short, Williams, & Christie, 1976; Sproull & Kiesler, 1991), the task (task determinism), or the social group mechanisms (sociological determinism, Harasim, 1993; Postmes, Spears, & Lea, 1998; Postmes, Spears, Sakhel, & de Groot, 2001; Riva, 2002; Short et al., 1976; Spears & Lea, 1992; Suler, 1996; Turoff, 1978).

Qualitative studies focused on the hermeneutics of CMC. Most of them came to conclusions that drastically differ from their quantitative counterparts, and depicted online interactions as warm and meaningful (for a meta-analysis see Walther, Anderson, & Park, 1994). The growing body of case studies (Hiltz, 1985) and ethnographies (Baym, 1993, 1998, 2000; Hine, 2000; Turkle, 1995, 1997) usually does not suffer from heavy methodological limitations, and many scholars attribute research findings to participant preferences and choices. Nevertheless, they very rarely factor the participants’ culture into the picture.

In tune with the most recent literature on online learning (Muleinburg & Berge, 2005; Mungania, 2003), Rice and Love (1987), Walther (1992, 1996), Herring (2001), and more recently Wood and Smith (2005) take a dispositional stand toward online exchange outcomes, according to which individuals adapt to a “lean” medium in order to build meaningful relationships. Again from online learning literature we know that three constructs are pivotal to engaging in online activities: (a) perceived control over one’s behavior (locus of control, Rotter, 1966); (b) belief in one’s capabilities to organize and execute the actions required to attain a goal (self-efficacy, Bandura, 1977); and (c) ability of having motivation from within (intrinsic motivation, Deci, 1975). These constructs are proven to be impacted by culture (Earley, Gibson, & Chen, 1999; Iyengar & Lepper, 1999; Santiago & Tarantino, 2002).

Despite the number of studies on CMC, neither the deterministic nor the dispositional stand fully account for how CMC determinants intertwine (Herring, 2001). Furthermore, neither addresses the impact of culture on CMC, while quantitative studies openly espouse determinism. However, if we postulate disposition to play a role in CMC production, no matter how small the role, the impact of culture on CMC can no longer be ignored.

**Linguistics and Semiotic of CMC**

Although the linguistic approach to e-mails provides a way to quantify spontaneous exchanges on an objective basis and can potentially bridge the gap between quantitative and qualitative findings (and their diverging interpretations), it has rarely been used (Lamerichs & Molder, 2003). Simplistically, social scientists and psychologists conceptualized e-mail as a form of written communication, which makes portability of findings into linguistics problematic (Herring, 2001). Even though linguistic scholars have repeatedly called e-mail a new genre (Ferrara, Brunner, & Whittemore, 1991) similar to what a transcript is to an oral exchange (Baron, 1998; Bordia, 1996; Lee, 1996; Herring, 1992, 2004), most studies involved caption and analysis of de-contextualized logs because the model of context is inadequate for e-mails (Herring, 1992, 2004; Jones, 2002).

In CMC, verbal cues are very strong. Communicators using a powerful language style are perceived as attractive, credible, and persuasive (Adkins & Brashers, 1995), and style is often more important than the actual exchange content (Pennebaker, Mehl, & Niederhoffer, 2003). Individuals tend to progressively match each other’s style and fine tune prepositions, articles, and other function words, regardless of the mutual liking or perceived quality of interaction (Niederhoffer &
Pronouns, words describing emotions, prepositions, and conjunctions have been associated with and can be used to predict a number of different outcomes (Newman, Pennebaker, Berry, & Richards, 2003). However, it is not known if these findings are also valid outside of North America.

CMC linguistics and discourse are unique because humans adapt to “lean” technologies by inventing paralinguistic expressions to reduce uncertainty and foster relationship formation (Herring, 2001; Walther, 1996; Walther & Tidwell, 1995), such as unconventional orthography to represent auditory information related to non-language sounds (MacKinnon, 1995), quotes from previous conversation to give the illusion of “taking turns” (Herring, 2001), emoticons to convey facial expressions (Herring, 2001; Rezabeck & Cochenour, 1994; Utz, 2000; Walther & D’Addario, 2001), and textual representations of physical actions to help contextualizing the conversation (Gumperz, 1982).

Counter-intuitively and despite the popularity of the “cues filtered out” approach, a number of non-verbal cues are available through CMC, such as chronemics (Hesse, Werner, & Altman, 1988; Liu & Ginther, 1999; Walther & Tidwell, 1995), which play a major role in impression formation (Liu, Ginther, & Zelhart, 2001; Walther, 1993), and relationship development (Hesse et al., 1988; Walther, 2002; Walther & Tidwell, 1995). Walther and Tidwell (1995) discovered that chronemics are contingent to e-mail orientation (i.e., task vs. social content). If task-oriented messages are sent at night, they are perceived as more dominant and “inherently” less social; the inverse is true for social messages. Also, quick replies to task-oriented messages and slow replies to social-oriented messages were rated more affectionate (Walther et al., 1994). This property can have a major impact on people living in different time zones.

Since the use and meaning of traditional verbal and nonverbal cues as well as paralanguage is modified by gender (Herring, 1992; Yates, 1993) and culture (Matsumoto, 2006), we can expect this to apply to e-mails as well.

Cross-Cultural Studies

There is no consensus on the definition of “culture,” because its meaning varies with context (Chase, Macfadyen, Reeder, & Roche, 2002). Early attempts conceptualized it as a set of acquired dispositions uncritically mutated from a homogeneous group (see Table 1 [Hofstede, 1983, 1991]; Table 2 [Hall, 1959, 1976]). More recently, Spencer-Oatey (2000) introduced an interpretive dimension according to which culture also influences interpretations of the meaning of others’ behaviors.

The main objection of Hofstede’s and Hall’s frameworks is the passiveness of both the individual toward the group and the group toward collective brainwashing. Furthermore, such a group is also homogenous and often restricted to ethnicity or nationality. Since both Hofstede’s and Hall’s works are based on data collected in managerial settings decades ago, they might no longer be valid (i.e., cultural “shift”), be valid only in business-related settings, or a combination of the two. Representing culture-related concepts in binary pairs may also miss the importance of cultural nuances, at least with regard to some constructs (Abdelnour-Nocera, 2002). Finally, other authors question en-bloc the idea of culture as a group phenomenon, operationalizing it as social negotiation of meaning (Scollon & Wong-Scollon, 1995). However, these objections might, in turn, just be a cultural reaction of researchers from an individualistic society who experience difficulty in accepting the power of social group conditioning on the individual.

Most studies on “culture” spanned three main thematic areas: (a) verbalizations of situations, feelings, values, norms, and thought patterns (Maletzke, 1996); (b) culturally sanctioned behaviors (Hall, 1960, 1966); and (c) interpretations of situations, such as conflict and collaboration.
### Table 1. Hofstede’s cultural dimensions

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Definition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Distance Index (PDI)</td>
<td>Degree of status difference among people</td>
<td>High PDI: inequalities of power/wealth are present within a society</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low PDI: differences between citizen’s power/wealth are undervalued</td>
</tr>
<tr>
<td>Individualism (IDV)</td>
<td>Degree to which individual (vs collective) achievement and interpersonal relationships are reinforced.</td>
<td>High IDV: individual is independent and placed above the group;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low IDV: the group and the individual are interdependent</td>
</tr>
<tr>
<td>Uncertainty Avoidance Index (UAI)</td>
<td>Level of tolerance for uncertainty and ambiguity within the society.</td>
<td>High UAI: low tolerance for uncertainty and ambiguity.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low UAI: less concern about ambiguity and uncertainty and more tolerance for a variety of opinions.</td>
</tr>
<tr>
<td>Long-Term Orientation (LTO)</td>
<td>Degree to which long-term devotion to traditional, forward thinking values is embraced.</td>
<td>High LTO: values of long-term commitments and respect for tradition highly regarded.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low LTO: long-term, traditional orientation and long-term traditions and commitments not very regarded.</td>
</tr>
<tr>
<td>Masculinity (MAS)</td>
<td>Degree of reinforcement of the traditional masculine work role model (achievement, control, power).</td>
<td>High MAS: high degree of gender differentiation.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low MAS: low level of differentiation and discrimination between genders.</td>
</tr>
</tbody>
</table>

### Table 2. Hall’s classical patterns

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Definition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context</td>
<td>High vs Low</td>
<td>High: less written/formal information, strong boundaries, situational knowledge;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low: more written/formal information, rule-oriented, task-centered, explicit and codified knowledge</td>
</tr>
<tr>
<td>Time</td>
<td>Monochronic vs Polychronic</td>
<td>Monochronic: Do one thing at a time, hard deadlines and plans, committed to the job, weak short-term ties, great respect for private property;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polychronic: do many things at once, highly distractible, committed to people more than time), change plans often, strong long-term ties, borrow and lend things often and easily</td>
</tr>
<tr>
<td>Space (proxemics)</td>
<td>High vs Low Territoriality</td>
<td>High: all the spaces around the individual are big;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low: all spaces around the individual are small.</td>
</tr>
</tbody>
</table>
(Hofstede, 1983). Concepts like degree of context expressed through communication (Hall, 1976), time (Hall, 1976), tasks/relationships focus (Laurent, 1983; Trompenaars & Hampden-Turner, 1997), attitudes toward authority, adherence to “traditional” gender roles, group/individual focus (Hofstede, 1991), universalism/particularism pair (Trompenaars & Hampden-Turner, 1997), ways to deal with conflict, decision making, and collaboration (Kim & Bonk, 2002; Trompenaars & Hampden-Turner, 1997) still hold validity, even in limited contexts and even in CMC settings (Ess & Sudweeks, 2005).

Interestingly, in a study of 39 languages, the ones from less individualistic cultures allowed for pronouns to be dropped from sentences (Kashima & Kashima, 1998), while Hinkel (1995) discovered that the use of modal verbs between native-speaker undergraduates and ESL students differs, and such differences do not change with linguistic proficiency and exposure to other socio-cultural constructs. These results are grounds to think that different metaphors stand behind different languages and are expressed differentially according to cultures.

Though cross-cultural studies suffer with conceptual and methodological limitations, intercultural studies have leapfrogged the cultural issue by switching from not considering culture to speaking of “intercultural communication through the Internet” without deepening the concept, validity, reliability of “intercultural” dimension, and its measurements (Chase et al., 2002; Macfadyen, Roche, & Doff, 2004; Scollon & Wong-Scollon, 1995). For this reason, old intuitions can still be used to build more rigorous frameworks.

If the adaptation to CMC is even partially dispositional, culture can influence our motivation in e-mail interaction engagement. If language and paralanguage are differentially shaped by culture, culture will also shape CMC. If our knowledge derives from a given set of mental schemes (i.e., metaphors), culture will also modulate how we enact those metaphors.

**MAIN FOCUS OF THE CHAPTER**

Different cultures adopt different metaphors within their languages, each one denoting what is important in the *weltanschauung* of that specific culture. Because every culture symbolizes its world in a different way, if we want to understand how metaphors are used and what we can do “with” them, we are left to make sense of metaphors. In this section I will first explore which metaphors are used in computer-related language. I will then build a model that puts such metaphors in relation to literature from the previous section and, finally, define a list of e-mail indicators to be used when testing the impact of culture in e-mail exchanges.

**Issues, Controversies, Problems**

Lakoff and Johnson (2003) adopted an interesting perspective about the role and impact metaphors have in structuring and shaping our comprehension of phenomena around us. They claim that humans understand and experience all experiences through the use of previous physical and social knowledge. The connection between previous knowledge and present topic resides in the embodied *metaphor*, rooted in physical experience without which there would be no abstract reasoning. Since we build thoughts and actions through metaphors, by analyzing which metaphors humans use for a given topic, we can understand which cognitive schemes their minds use in representing, framing, and understanding the phenomenon, its consequences on the representation itself, and possible “actions” around that metaphor.

Possible claims against Lakoff and Johnson’s theory focus on two concepts: (a) thought is complex and unconscious for the most part, and (b) abstract concepts are mostly metaphorical. Murphy (1996, 1997) explained technical (psycholinguistics) objections, while Gibbs summarized his stand in favor of the embodied metaphor concept (2003). From a behavioral perspective, it can be
argued that we learn language from the environment, as a response to stimuli, and therefore it does not represent a complex unconscious scheme but just an adaptive reaction. Such reaction would however fail to explain why different metaphors are generated by the same person in a given situation. Alternatively, we could argue with Gumperz (1982) that metaphors are expressions of culture rather than manifestations of cognitive schemes. Such objection, however, would not explain why similar metaphors originate in different situations, environments, or cultures. Finally, a constructivist would object that objective meaning does not exist per se, but only as a negotiation process, therefore analyzing language from a metaphorical perspective de-contextualizes it.

The nouns commonly used when referring to the Internet refer to it as a “space” with boundaries, simultaneously away from where we are (“virtual”) yet earthly present (“world”). Such space is partitioned into smaller ones (“chat rooms,” “channels,” “message boards”) which mirror, in name and behavioral conventions, their “real” counterparts (greeting other participants when joining, introducing oneself, saying “bye” while leaving). Members already present in such space but unwilling to engage in a dialogue (lurkers) exist unobserved at its margins and posts, and messages are held together through “threads,” which implicitly suggest: (a) the presence of discrete separators within public spaces, and (b) the virtual community to be a spatial object with a definite dimensionality brought about by the degree of willingness with which its members interact. Microsoft espoused this metaphor and gave it an adventurous gist through names such as “Explorer,” “Windows,” and “Outlook,” inviting the user to “come out” into a broader space and “take a look.”

Verbs describing Internet activities configure the “Internet is an informal journey” metaphor. Users “go to” one page from another, meanwhile they casually “chat” with whom they meet, superfluously “browsing” information and “surfing” Web content.

The metaphors behind computers and their usage and communication in general go against the “negative” view of computers and CMC that emerged from the quantitative literature review. Schwartz (as cited in Lakoff, 2006) noted how the metaphor behind general machine functioning is “machines are people,” in which the degree of functioning expresses their health while the computational power is seen as “thinking.” Interestingly, the interaction between people and their computers also follows a human-to-human metaphor (Nass & Steuer, 1993). In the case of computer applications, they “go crazy,” “freeze,” and have to be “killed,” “terminated,” or “shut down.” They do not just have a “life,” but computer difficulties in handling the user’s requests are seen as emotional trouble. This last trait conflicts with the metaphor of computers and their parts as working aids for their owners.

Espenson (as cited in Lakoff, 2006) noted how the metaphor behind communication is “communication is linguistic communication.” Text-based or spoken communication features are considered more meaningful than nonverbal cues (“His eyes said yes.” “His look spoke volumes.” “I can read him like a book.” “His gestures underlined what he was saying.”).

Because some have advocated language as the source of cyberspace (Cicognani, 1998), and CMC has been represented in different cultures and languages through a relatively uniform set of metaphors, exploring these very same metaphors may teach us which tools to use when studying the correlations between culture and CMC.

Solutions and Recommendations

If the cognitive schemes applying to CMC result from the interplay of these metaphors, we can expect them to be critically enacted depending on the cultural framework of the interaction participants.
that is, depending on the relative importance and acceptance each metaphor has in a given culture. Since both face-to-face (Kramsch & Anderson, 1999) and online communication patterns (Chase et al., 2002) do vary among cultures, variations in qualitative CMC characteristics might also exist.

The Internet is a space. Proxemics are not limited to physical ones (Riva & Galimberti, 1998) and hold validity also for avatars in virtual environments (Jeffrey & Mark, 1998). For these reasons, some behaviors enacted through e-mails and some specific characteristics of the e-mail can be perceived as an intrusion into the receiver's space and be accepted (or condoned) in situations that vary with cultures. I therefore posit that initiating an unsolicited exchange, the e-mail length, and the e-mail rate are space domain indicators (see Table 3). These indicators might characterize cultures in which a hard work ethic is paramount, and which either value privacy, time, and wide individual spaces (such as North America) or shyness, social order, and conformity (such as China and Japan).

The Internet is informal. Culture affects conversation shape and content, as well as degree of technology adoption (Buragga, 2002; Phillips, Calantone, & Lee, 1994). Novelty-seeking, innovative behaviors and authority and rule-probing/questioning through e-mails could be equally culturally sensitive. For these reasons, I posit the formality domain indicators (see Table 4) to be dependent on culturally sanctioned behaviors with regard to public expression of emotional content and to encompass: (a) e-mail orientation (social/task-related content), (b) number of emoticons

<table>
<thead>
<tr>
<th>Table 3. Space domain indicators (virtual proxemics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicator</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Initiation</td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td>Rate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4. Formality domain indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicator</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Hierarchy</td>
</tr>
<tr>
<td>Tradition</td>
</tr>
<tr>
<td>Cues</td>
</tr>
<tr>
<td>Emoticons</td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td>Linearity</td>
</tr>
<tr>
<td>Orientation</td>
</tr>
<tr>
<td>SelfDisclosure</td>
</tr>
</tbody>
</table>
per e-mail, (c) number of other paralinguistic cues (such as emoting, quoting, unconventional orthography, and textual representation of physical actions), and (d) degree of self-disclosure. Other such indicators reflect important culture-related attitudes toward (e) hierarchies and (f) traditions, while (g) e-mail linearity (defined as logical, low-context, step-by-step string of thoughts), and (h) e-mail length reflect the attitude toward reflexivity and extroversion, respectively. Such indicators might be particularly descriptive of cultures that value emotional self-disclosure, extroversion, multi-tasking, traditions, and hierarchies (such as the Mediterranean cultures).

The Internet is a journey. In monochronic cultures, time is compartmentalized, while in low-context ones, competition and order are highly valued. In monochronic low-context cultures, work is conceptualized as a “serious” activity during which no derailments and few distractions are allowed, while journeys are meaningful if they have a goal. I posit that journey and working domain indicators (see Table 5) are: (a) daily differences in e-mail orientation (either polychronic switching/multi-tasking between emotional and task-related e-mails, or monochronic string of work-related e-mails interspersed with occasional personal e-mails or remarks), (b) e-mail linearity

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronemics</td>
<td>Walther’s chronemics</td>
</tr>
<tr>
<td>Rate</td>
<td>Email rate</td>
</tr>
<tr>
<td>ΔLength</td>
<td>Daily differences in email length (variations in email word count)</td>
</tr>
<tr>
<td>ΔLinearity</td>
<td>Daily differences in email linearity (variations in how the context is expressed)</td>
</tr>
<tr>
<td>ΔRate</td>
<td>Daily differences in email rate (variations in email daily number)</td>
</tr>
<tr>
<td>ΔOrientation</td>
<td>Daily differences in email orientation (switching, multitasking, or any variation of subject of the task vs. relationship)</td>
</tr>
</tbody>
</table>

Table 6. Text domain indicators

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarifications</td>
<td>Presence/absence of request for clarifications</td>
</tr>
<tr>
<td>Linearity</td>
<td>Email linearity (step-by-step vs. loose)</td>
</tr>
<tr>
<td>Orientation</td>
<td>Task vs. Relationship</td>
</tr>
</tbody>
</table>
(attitude toward context), (c) e-mail rate (number of tasks and kind of topics the sender has been engaged in daily), (d) e-mail length (extraversion and/or low-context culture), and (e) Walther and Tidwell’s chronemics (to be analyzed together with e-mail orientation). This list of indicators could be useful in exploring polychronic low-context cultures, such as Australia.

Communication is linguistic communication. Low-context cultures rely more on written information, are more verbally explicit and less emotional, and ask for details whenever they do not possess all the information they desire to have (Gill, 1998). When looking at data from Europe on the differential use of text-based CMC, studies reveal higher e-mail usage in northern (low-context) countries and higher chat room usage in southern (high-context) nations (Greenspan, 2002). This can be explained by positing that in high-context cultures, e-mail is less frequent because asynchronous communication is perceived as de facto unable to convey “all the context” or not fast enough. For these reasons I posit that (a) requests for clarifications, (b) e-mail orientation, and (c) e-mail linearity are text domain indicators (see Table 6). This indicator cluster might be an excellent fit to explore the low-context/high-context dimension.

Machines are people. Cultures that value long-term relationships and high-context cultures encourage people to behave traditionally and collaboratively toward the group (Kim & Bonk, 2002), and attend to long-term relationships. In individualistic cultures, abstract nouns and adjectives predominate, as opposed to interpersonal concrete verbs within collectivist cultures (Semin, Gorts, Nandram, & Semin-Goossens, 2002). Furthermore, collectivist cultures approach online decision making with different values (Rahmati, 2000; Thanasankit & Corbitt, 2000), while individualistic cultures have a higher rate of critical exchange (Reining & Mejias, 2004). For these reasons, I posit that (a) the attitude toward collaboration, (b) traditions, and (c) individuals, as well as (d) the emotional e-mail lexicon type (abstract nouns/adjectives vs. interpersonal concrete verbs), (e) the existence of long-term exchanges with one or more in-groups, and (f) the number of critical exchanges, are people domain indicators (see Table 7). These indicators can explore the individual/group dimension.

Looking at possible e-mail cultural aspects through the lenses of the five metaphors identified in the previous section allowed the author to identify 22 putative indicators to be put through validation in further studies.

---

**Table 7. People domain indicators**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaboration</td>
<td>Attitude toward collaboration</td>
</tr>
<tr>
<td>Hierarchy</td>
<td>Attitudes toward hierarchies</td>
</tr>
<tr>
<td>Tradition</td>
<td>Attitudes toward traditions</td>
</tr>
<tr>
<td>Critical Exchanges</td>
<td>Messages expressing critical viewpoints</td>
</tr>
<tr>
<td>Lexicon</td>
<td>Emotional email lexicon</td>
</tr>
<tr>
<td>LongTerm</td>
<td>Existence of long-term exchanges with one or more in-groups</td>
</tr>
</tbody>
</table>
FUTURE TRENDS

Possible Scenarios

Knowing how different cultures perceive e-mails and their characteristics will be a useful tool in various settings. For-profit and non-profit corporate managers alike could use these findings to align e-mail content, format, and policy to their organizational values, and informally manage enculturation, employee loyalty, and adherence to corporate culture. Human resource development managers and online educators could foster cross-cultural understandings and proper management of distant colleagues/partners, while also increasing knowledge sharing and collaboration whenever desired by tailoring their CMC to the audience’s culture, motivation, and desired workplace or educational outcomes. Senior managers and supervisors could facilitate the social component of merging and outsourcing, by planning online hiring procedures and online job assignments, as well as disseminating electronic memos, news, and training manuals in a culturally sensitive way. Online trainers and educators could exploit online social learning by tailoring the format and content of e-mail exercises and interactions, which would allow them to differentially involve students in full respect with their cultural requirements, while still targeting desired learning objectives and outcomes. Professionals like health providers, academics, knowledge managers, and software developers would benefit worldwide from a blueprint for culturally sensitive interactions with peers for informal training, knowledge generation, and referral. Finally, job hunters and potential students could learn how to write personal electronic materials (i.e., electronic college applications, electronic curriculum vitae, cover e-mails) designed to appeal to like-minded individuals.

Limitations

The embodied metaphor’s perspective main problem is that culture can influence CMC prior to the moment in which a metaphor is enacted or interpreted, since the attitude toward using technology is related to specific cultural aspects (Buragga, 2002; Gill, 1998; Hermeking, 2005; Van Belle & Stander, 2002). However, when analyzing spontaneously occurring e-mail exchanges, this objection seems less poignant because CMC users either belong to a technology-friendly culture or have already overcome their own cultural bias toward technology. The objection is still valid whenever e-mail usage is mandated, and is a serious pitfall of some research study designs if unaddressed. It remains to be assessed if spontaneous e-mail usage is an indicator of relative easiness in overcoming cultural biases in general.

The cross-cultural perspective per se is another possible limitation. Language on the Web “becomes” culture (Kramisch & Anderson, 1999), so CMC might benefit from an inter-cultural perspective instead (Bennet, 1993; Chase et al., 2002; Ma, 1996; Macfadyen et al., 2004). On the other hand, to the newcomer to the Internet and to individuals who have not yet criticized their culture, intercultural perspectives may paradoxically foster cultural insensitivity. The Internet itself may also generate glocalization in dealing with cultural aspects of CMC, especially in public fora where adherence to norms is strongly encouraged.

Although all five domains are developed (see Figure 1) and possess both unique and shared indicators, many indicators are not specific to a category, presenting significant overlapping.

That might reflect a poor metaphor choice as well as nuanced interconnections among metaphor dimensions, or frank interdependence. The best characterized domains are “People” (six indicators, four unique) and “Formality” (nine
Figure 1. Relationship among domain indicators

![Diagram of relationship among domain indicators]

Table 8. List of indicators

<table>
<thead>
<tr>
<th>Qualitative</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initiation</td>
<td>Length</td>
</tr>
<tr>
<td>Hierarchy</td>
<td>ΔLength</td>
</tr>
<tr>
<td>Tradition</td>
<td>Cues</td>
</tr>
<tr>
<td>Collaboration</td>
<td>Linearity</td>
</tr>
<tr>
<td>Clarifications</td>
<td>ΔLinearity</td>
</tr>
<tr>
<td>Orientation</td>
<td>Rate</td>
</tr>
<tr>
<td>ΔOrientation</td>
<td>ΔRate</td>
</tr>
<tr>
<td>SelfDisclosure</td>
<td>Emoticons</td>
</tr>
<tr>
<td>Critical Exchanges</td>
<td>Lexicon</td>
</tr>
<tr>
<td>LongTerm</td>
<td>Chronemics</td>
</tr>
</tbody>
</table>
indicators, three unique). The latter appears to be the mainstay of the model, also allowing for use as indicator subset for gross model validation. The 22 indicators are evenly split between methodological categories (11 quantitative, 11 qualitative—see Table 8).

Testing of the Model

In order to build an objective theory that matches sample accuracy, it is mandatory to match the complexity of the data set with the one of the model. Among available methods, action science (see Argyris, Putnam, & Smith, 1985) appears to be the most fit because it explores events that: (a) occur spontaneously, (b) are composed of (or can be subdivided into) discrete quantifiable measurements, and (c) are not adequately explorable by an artificial reproduction of their core conditions. The paucity of studies on CMC in “natural” settings and the impossibility to fake or eliminate “culture” point toward the use of observation of “spontaneous” interactions.

Previous studies on CMC belonged to either the positivistic model, where the researcher’s epistemology and ontology are both objective, or the postmodern framework, in which the researcher’s epistemology and ontology are both subjective. Because of the conditions in which both CMC and “culture” occur, the objective ontology and subjective epistemology of action science offer a fresh perspective. Action science also offers excellent guarantees about being the least biased method in these settings, since it does not make a priori hypotheses and does not embed interaction design into the study.

Finally, action science also shares its ontology with Lakoff and Johnson’s (2003) model. The parallel between Argyris’ cycle (observation, model building, testing, and action; Argyris et al., 1985) and metaphors as products of the interplay between objectivism and subjectivism is very strong. Metaphors can act as common ground between one cycle stage and the next one, or between cycles.

Researchers bring into the study their culturally susceptible biases, behaviors, and interpretations (Cray & Mallory, 1998), so a great deal of attention should be paid to study design criteria. Viable strategies are: (a) choosing an intercultural study design, and/or (b) employing multiple teams from different cultures observing, analyzing, and interpreting the same set of data so to either shrink the “cultural divide” or make it apparent.

Data selection has to be carefully done. Researchers should collect spontaneously occurring exchanges before enrolling participants-to-be, to prevent a change in their spontaneous motivation to participate into the exchange from affecting the linguistic of the data under study. Data can be gathered from public fora, Yahoogroups, Google groups, message boards, and listserves, as well as from any person of legal age and capacity to read, understand, and sign an informed consent privacy disclaimer describing type and scope of information data collection. Researchers should then contact and test participants via a standardized test, exploring as many independent cultural dimensions as possible, to determine which values participants do individually espouse. Researchers should code and analyze data before formal participant enrollment or testing, and in any case, without knowing the writer score while coding, scoring, or analyzing the message, so to avoid a source of interpretive bias. The obtained data could undergo a cluster analysis to see if taxonomy similar to the original model is produced.

Future Research

Even after successful testing of this framework, there will be further questions to address design and content of future research studies.

Among the contents to be addressed, a first big bulk of questions to explore relate to the difference between CMC in dyads and in groups,
and the interplay between these differences and culture. These questions correspond to the study of possible intersections between the people and text domains (see Figure 1). How does the presence of more than two persons (group) change the dynamics of CMC exchanges? Are cultural expectations related to participation in group conversations? Are there other norms regarding public interactions that change with culture (for example, is the concept of “face” valid online as well)?

Another big set of questions to research lies in the interplay between constructs such as motivation, self-efficacy, and locus of control, with CMC via the cultural and cognitive metaphors applied to those constructs themselves. These questions would explore possible intersections between the journey/working and people domains (see Figure 1). Are motivation, self-efficacy, and locus of control linked to any specific cultural dimension? How does it impact on technology adaptation? How does it impact on the tendency to transcend one’s culture?

Finally, a further set of questions explores the feedback that CMC may have on culture, which would refine the relationship between the people and formality domains (see Figure 1). Which kind of context is available in different cultures as far as e-mail exchanges are concerned? When, why, and how does the perception of cultural norms about e-mail change how different cultures “do e-mails”? What is the impact that signatures and pseudonyms have on CMC exchanges?

This chapter could provide for a set of CMC metrics that could quantify the impact culture has on CMC, thereby changing the current approach to CMC. Cultural inclusiveness, embedding a representation into the interaction, is not customizable enough for the needs of users, even if it claims to be so. It has to be pointed out how “cultural inclusiveness” is a concept germane to political correctness, which in turn is deeply rooted into the American culture and would therefore be yet another “symptom” of cultural imperialism.

Intercultural sensitivity, on the other hand, recognizes and acknowledges different cultures in a framework in which one’s culture is examined critically, while the other person is evaluated from within his/her culture (Bennett, 1993).

If this paradigm will be validated, it might really impact CMC research. The use of metaphors to make sense of CMC has never been attempted before and could be extended toward other forms of CMC such as IM. Findings from these studies could change the policies on the usage of CMC, with particular attention to workplace settings, and that in turn could change people’s behavior and perceptions about CMC.

CONCLUSION

Despite the increasing use of asynchronous text-based CMC in cross-cultural settings, most of what we know about it comes from studies that did not take culture into any account. As a general rule, CMC research has either opted for “laboratory” conditions in the attempt to quantify measurable behavior, or used qualitative methods in the attempt to capture the complexity of field conditions, but results have been channeled into determinism of some sort. The approach this chapter illustrates is a promising one because it refuses to take “absolute” sides, yet addresses qualitative and quantitative issues in the attempt to match complexity of the problem with complexity of the method of study.

The way culture has been conceptualized and studied has shown to be outdated not merely because it dates back to the 1960s: methodological concerns arise toward the generalization of findings from special populations to the whole humankind. However, some of those concepts and findings can still be used to frame actual research. This chapter uses those findings as categories, in a way as “tags” to start from for classifying events and their properties, and then reframe them to explain everyday phenomena.
Building a model based on the metaphors that most frequently are associated with CMC and looking at those metaphors through frameworks accounting for cultural differences allows the author to find a list of qualitative and quantitative behavioral and linguistic metrics (see Table 8) potentially able to depict how culture impacts on CMC exchanges, going beyond the stands that have dominated the field so far. In fact, such metrics cut along disciplines and are related to one another in a way that also allows predictions about their change depending on starting conditions and changes that can occur in them.

This chapter mixes those overlapping disciplines that have most frequently studied CMC into a consistent framework that traces the path for future research. Action science is described as a way to conjugate metrics grounded in an interdisciplinary framework, while implementation details about future research suggest area of improvement.

REFERENCES


Liu, Y. (2002). What does research say about the nature of computer-mediated communica-
Linguistics of Computer-Mediated Communication


Postmes, T., Spears, R., & Lea, M. (1998) Breaching or building social boundaries. Side-effects of


Chapter 5.11
Impression Formation in Computer-Mediated Communication and Making a Good (Virtual) Impression

Jamie S. Switzer
Colorado State University, USA

ABSTRACT
In face-to-face interactions, people generally form impressions by focusing on a variety of nonverbal cues. Increasingly, however, people are communicating virtually and forming impressions based on mediated interactions. In an online environment, the range of nonverbal cues that normally aid in impression formation is drastically narrowed. In the absence of these nonverbal cues, forming impressions via computer-mediated communication places a greater emphasis on verbal (text-based) and linguistic cues. This chapter offers strategies to ensure virtual workers make a good impression on their clients and colleagues when interacting online.

INTRODUCTION
As the saying goes, you never get a second chance to make a good first impression. This is especially true when working virtually, where impressions are formed via computer-mediated communication (CMC). As Wallace observes, “Increasingly…the online persona is playing a larger role in first impressions as people rely on email, websites, and discussion forums more for the first contact, and the phone call, letter, or face-to-face meetings less” (1999, p. 14).

The varieties of nonverbal cues that normally aid in impression formation do not exist in an online environment. In the absence of these nonverbal cues, forming impressions via CMC places a greater emphasis on verbal (text-based) and linguistic cues, as well as depending more
Impression formation in Computer-Mediated Communication and Making a Good (Virtual) Impression

upon social cues such as shared schema, context, and stereotypes. Indeed, as Tannis and Postmes observe, “communications over the Internet are all but free from influences of the social, the cognitive, and the physical” (2003, p. 692).

It is critical, then, that people working virtually understand how impressions are formed in an online environment and what types of cues aid in forming those impressions, so virtual workers can manage their own online behavior in such a manner that allows others to form an accurate impression.

BACKGROUND

Impression formation is a significant characteristic of communication and a fundamental social-psychological process (Walther, 1993; Liu, Ginther, & Zelhart, 2002). Asch (1946) and Goffman (1959) are generally associated with the earliest scholarly research into impression formation. Regardless of the nature of an interpersonal interaction, as humans “we seem to exit most of our social encounters with some general impression of the other person’s characteristics and dispositions” (Hancock & Dunham, 2001, p. 325). It is simply human nature to form impressions of those around us with whom we communicate in a variety of ways and with differing motivations and goals.

Early scholarship into impression formation emphasized how traditional cues and sources of information identified as important in face-to-face interactions were reduced or eliminated in a CMC environment (Short, Williams & Christie, 1976; Kiesler, Siegel & McGuire, 1984; Sproull & Kiesler, 1986). These theories, collectively termed the “cues filtered out” approach (Culnan & Markus, 1987), concluded that a lack of nonverbal cues prevented people from forming impressions in CMC.

However, a growing body of empirical research has since contested the findings of the cues filtered out theories, and instead has shown that people compensate for the lack of nonverbal cues in a variety of ways and do indeed form well-developed impressions in mediated environments (see Sherman, 2001, p. 54). One of the more recent theoretical models that is key to understanding impression formation in CMC is social information processing theory (SIP).

SIP (Walther, 1992) posits that in the absence of nonverbal cues, people adapt. They are motivated to use whatever information they have available in a particular medium to provide cues to assist in impression formation. According to SIP, while certain nonverbal cues are missing in CMC, other cues needed to form impressions are still exchanged during an interaction. Because of the nature of CMC, the process takes more time and impressions are formed over a more extended period than in face-to-face interactions.

IMPRESSION FORMATION CUES IN CMC

As predicted by SIP, human beings rely on specific types of cues inherent in CMC to form impressions. Categories, stereotypes, schemas, cultural background and preconceived biases all influence the formation of impressions. Studies have shown that the impact of these social cues is considerable in impression formation (Tanis & Postmes, 2003). In CMC, the number of social cues is reduced, but that “does not point to a reduction in the social context of the CMC” (Spears & Lea, 1992, p. 324).

Social stereotypes and exemplars are types of metonymic models (models in which one member of a category is used to understand the membership as a whole) that are commonly utilized when forming impressions in an online environment (Jacobson, 1999). A stereotype is a model where a society or culture recognizes characteristics of an individual or group of people as representing an entire category. For example, some athletes often make headlines by engaging in unruly
and unlawful behavior. Thus athletes are often categorized as being rowdy and uncontrollable, which is more likely the exception instead of the rule. There are many, many athletes who devote their time, energy, and money to charitable causes, who are dedicated to their families, and who do not break the law. However, athletes as a whole are often stereotyped according to this commonly held belief that athletes are unmanageable.

An exemplar is a specific individual that a person has encountered, who is then taken as a representative of others who are thought to be in that same category. However, different people have different experiences with different people who serve as exemplars. Using the previous example of athletes, if someone met an athlete at a charity function who was talking and interacting with the crowd, happily signing autographs and taking photos with children, the inference might be that this particular athlete was a nice person. Based on this one encounter, that athlete is an exemplar for all athletes, who must therefore all be nice people. That is certainly not the case, particularly in a few specific sports.

Impression formation in CMC, as in face-to-face interactions, is extremely dependent upon social cues such as stereotypes, along with the context of the communication and shared schema. People tend to fill in their gaps of knowledge about another person with typifications based on stereotypes and exemplars, a concept defined as “idealization” (Goffman, 1959). Within that framework, people need to seek as much information as possible to form an impression.

Impression formation refers to the “interpersonal processes by which people employ all available information and make general judgments of others’ personality characteristics” (Liu, Ginther, & Zelhart, 2002, p. 73). In a face-to-face environment, people rely on nonverbal cues like vocal patterns (such as tone or accent), linguistic markers (such as vocabulary and grammar), body posture, gestures and eye contact to glean information. The words that are actually spoken are not as influential when forming an impression (Wallace, 1999). But those nonverbal cues are missing when working virtually—the range of cues available to participants in the interaction is narrowed in mediated interactions (Lyon, 2002).

In an online environment the impressions people form and the judgments they make are based solely on the information available to them. People create a perception of others to the extent that they extensively process available information (Leyens & Corneille, 1999) within the context of their own biases and culture. However, in an online environment, that information is extremely limited and is usually conveyed mainly via keystrokes, making the keyboard an “unfamiliar and awkward impression-making device” (Wallace, 1999, p. 28). Yet there are many ways—some subtle, some obvious—that information is communicated in CMC.

E-mail addresses and domains can convey a considerable amount of information about someone’s personality, character, and status. “John. Smith@mycompany.com” gives the impression of professionalism; “beachluvr@hotmail.com” does not. An e-mail coming from a “dot edu” domain indicates that the person is in some way involved in education, which implies a more intellectual status than a “dot com” domain. However, there is ambiguity within a message received from a “dot edu” domain. Is the sender a teacher or professor? A student? An athletic coach? A maintenance worker? To form an accurate impression of the sender of the message, the receiver must rely on additional textual and linguistic cues.

Linguistic style in discourse has a significant impact on impression formation. Users of powerful language styles are perceived as competent, credible, attractive, and persuasive (Adkins & Brashers, 1995). Likewise, in CMC, additional cues such as language use, word choice, vocabulary, sentence structure, and spelling also provide information about a person.

For instance, in the previous “dot edu” example, if an e-mail included a number of references to a
complicated scientific process and long, unpronounceable words, one may conclude the sender was a professor. If the message contained words such as “dude” and “awesome” within short, incomplete sentences with informal or even incorrect spelling (such as “kewl” or “howRU”), the impression is that the sender is an undergraduate student.

Obviously there can be ambiguity within language styles. A professor might very well use the word “awesome” and an undergraduate be well versed in the language of science. In general, however, our shared social schemas lead to forming those particular impressions. According to the stereotype, people believe professors use large words when communicating, and stereotypical undergraduate students use slang.

Likewise, the presence of spelling errors in an online interaction can have multiple meanings. Was the sender just in a hurry and did not closely proofread the e-mail? The impression one might form is that the sender is too busy to be bothered with proper spelling, meaning perhaps it will be difficult to collaborate virtually with this person because they are overworked. Does the sender truly not know how to spell? The impression then might be that the sender would not be a viable partner in a virtual business collaboration.

Vocabulary, word choice, and spelling can also provide information about nationality and/or geographic location. If one of the participants in an online interaction used the word “boot” when referring to the back end of a car, the impression is that the person is from England and not the U.S. (where the term would be “trunk”). Likewise with spelling: the word is spelled “color” in the U.S. and “colour” in England.

Even within countries, linguistic cues can provide geographic information. The term “y’all” is typically associated with the southern region of the U.S. If a message is received containing that term, the impression is that the sender is from the South, which in turn conjures up a host of other stereotypical impressions (usually unflattering) of people who live in the southern U.S.

Information can be communicated in CMC using paralinguistic cues, which can also assist with impression formation. Typographical marks and other textual features such as exclamation points (excitement), ellipses (trailing thought), emoticons (smiley faces), and use of capital letters (SHOUTING) have no lexical meaning per se. The meaning of paralinguistic cues is “dependent on the group or individual context that is pre-established for the communication” (Lea & Spears, 1992, p. 321); to understand them, the meanings must be socially shared among all participants in the interaction. Based on the context, impressions formed using paralinguistic cues can be positive or negative.

Another nonverbal cue utilized when forming impressions in CMC is the use of time, known as chronemics (Walther & Tidwell, 1995). The timing of message sending and receiving, as well as the frequency and duration of online interactions (Liu, Ginther, & Zelhart, 2002), influences impression development in CMC and provides information that can be interpreted in a variety of ways, particularly in the context of a work environment. For example, if a person sends an e-mail late at night, the impression may be that sender is a hard worker who spends many long hours at the computer. If a receiver waits a long time to respond to a message, it may signify superiority of the receiver or a perceived lack of status on the part of the sender. An immediate response implies priority and importance.

However, like all other nonverbal or paralinguistic cues, chronemic cues have their greatest influence in the context of other cues and relationships (Walther & Tidwell, 1995). It may be that the late-night e-mailer works from home and is most productive at night and sleeps during normal business hours. Therefore, the impression of the person as a hard worker may not be correct and other cues are needed to form an accurate impression.
Impression management is the process by which individuals attempt to control others’ perceptions of them, motivated by the desire for social acceptance and relationship development and maintenance (Becker & Stamp, 2005). People spend a great deal of time on impression management, working to regulate what information is known and not known by others about oneself in order to manage the impressions others have of them (O’Sullivan, 2000). Impression management theory posits people are motivated to control the impressions they make on others (Schlenker, 1980), regardless of whether or not the interaction occurs face-to-face or virtually.

Studies by Asch (1946) showed long ago that people tend to leap to conclusions very fast, with few cues to guide them, when forming first impressions. First impressions, however, are “notoriously susceptible to misperception” (Wallace, 1999, p. 15) because people base their impressions on minimal cues and all other assumptions flow from that initial conclusion, regardless of its accuracy.

Additionally, initial impressions often lead people to behave toward another person in certain ways (Sherman, 2001). Sherman gives the example of perceiving someone to be friendly and therefore interacting with him or her in an open and welcoming manner; the other person, then, may respond in a manner even more friendly than perhaps they would normally. This is referred to as the “snowball effect” (Gilbert, 1995), in which a first impression can initiate interactions that gather momentum and become harder to alter and correct if they are indeed inaccurate. First impressions are important because people do not like to admit mistakes. This can lead to confirmation bias, where people ignore evidence that may contradict their first impressions and even actively seek information to confirm the original impression, even though it was incorrect (Wallace, 1999).

Self-presentation is a technique used in impression management. Senders of messages can take advantage of the nature of CMC to carefully craft messages and deliberately enhance the representation of themselves by optimizing self-presentation (Walther, 1996). By engaging in selective self-presentation, people can control what information is revealed during an interaction, thereby influencing what cues are available for the receivers of the message to form impressions.

In some ways, impression management is easier to control in CMC because there are no accidental or unintended nonverbal behaviors or physical cues (Walther, Slovacek, & Tidwell, 2001). Additionally, visual cues may not be present, so there is no preconceived stereotypical perception that can prejudice the interaction.

In other ways, however, impression management in CMC is more difficult than in a face-to-face environment. Research has shown that physical appearance is extremely important in the formation of first impressions (Hatfield & Sprecher, 1986); in most cases that cue is absent in CMC. Words can be construed with different meanings than originally intended in the absence of nonverbal cues. Text-based communications can be wholly misinterpreted by the receiver of the message because there are no additional social cues present to aid in understanding the message more accurately. It is important, then, that virtual workers manage their online behavior in such a manner that others have the information they need to form an accurate impression.
STRATEGIES FOR MANAGING IMPRESSIONS IN CMC

“The desire to form impressions of other people and to manage our own impressions in social settings...does not disappear just because we now do these things on the Internet” (Wallace, 1999, p. 36). Virtual workers must be acutely aware of the nature of CMC in order to make a good virtual impression on their clients and colleagues when interacting online. As predicted by SIP, people form impressions based upon one’s online behavior, conveyed using verbal (text-based) and linguistic cues in the context of cultural background and personal biases. In the absence of any “reality checks” (Jacobson, 1999) such as nonverbal cues or visuals, virtual workers must employ a variety of strategies to carefully manage their interactions to gain social acceptance as well as develop and maintain viable relationships.

CMC users employ a variety of uncertainty reduction cues to “give and gain impressions of one another” (Walther, Loh, & Granka, 2005, p. 40). According to Uncertainty Reduction Theory (URT), the exchange and collection of information allows people to predict the attitudes and behaviors of others (Berger & Calabrese, 1975). The more information that is gathered about a person, the more uncertainty toward that person is reduced (Tidwell & Walther, 2002).

URT posits that individuals utilize three different information-gathering strategies as a means of reducing uncertainty: passive, active, and interactive (Tidwell & Walther, 2002). Passive strategies involve the unobtrusive observation of a person to garner information. Active strategies involve targeted efforts to collect information about someone, but without that person’s knowledge. Interactive strategies for reducing uncertainty entail direct exchanges among the people involved.

As predicted by URT, people are going to be actively seeking information with which to form an impression of others. In CMC, it is difficult to “observe” a person’s behavior, but it is very easy to seek information about someone either actively (by Googling them, for example) or interactively via direct dialogue. Therefore, virtual workers must manage their own behavior in such a way that allows others to form an accurate impression of them in a mediated environment. Below are some strategies to ensure that virtual workers make a good impression on their clients and colleagues when interacting online. While these uncertainty reduction strategies may seem relatively simplistic and straightforward, they are the most accessible and effective to employ in CMC.

Full Disclosure

One way to reduce uncertainty and prevent idealization of senders and receivers of messages is to voluntarily disclose as much relevant information as necessary in the context of the virtual workplace and the task at hand. The sharing and acquisition of information is a proactive strategy that can significantly reduce any uncertainty present due to the lack of nonverbal cues in CMC (Tidwell & Walther, 2002).

Writing self-descriptions is one way to share information. Introducing oneself to virtual colleagues and customers, providing biographical details relevant to the establishment and development of the relationship (such as years of experience in the field), is an excellent strategy for managing impression formation. People have a tendency to increase the amount of information shared (self-disclosure) when interacting via CMC (Joinson, 2001), thus providing additional cues to aid in impression formation. Research has shown that even minimal cues, such as a few biographical details, can have a drastic impact on the quality of impressions people form of one another (Tanis & Postmes, 2003).

Be Honest

There is considerable uncertainty about the accuracy of information in CMC, particularly with
Impression Formation in Computer-Mediated Communication and Making a Good (Virtual) Impression

respect to self-presentation and how closely that matches how people really are (Walther & Parks, 2002). Honesty is critical when interacting virtually. It can be tempting when managing impressions to not be absolutely forthright, particularly if there is no chance the virtual working relationship will ever move to a face-to-face environment.

People employ a variety of information-seeking strategies beyond direct online interaction, such as acquiring information from third party sources without the other person’s knowledge or using search engines to discover information that is available on the Internet (Ramirez, Walther, Burgoon, & Sunnafrank, 2002). So it is likely that if a person is not honest when communicating virtually, that deception will be discovered. Likewise if an initial impression is inaccurate when it is articulated, if left uncorrected, the “snowball effect” discussed previously can occur.

Research suggests, however, that people do not actively deceive others when trying to form relationships and manage the impression others develop in an online interaction (Albright, 2001). Even “to reveal oneself as clearly and honestly as possible requires individuals to select from a repertoire of self-presentational activities to convey themselves accurately,” suggesting that self-presentation is not necessarily manipulative or deceptive, but “should be viewed as a broader concept that includes efforts to present oneself accurately” (O’Sullivan, 2000, p. 406).

Additionally, people who know that the virtual relationship will eventually lead to a face-to-face meeting tend to be more open and honest in the self-disclosure process (Gibbs, Ellison, & Heino, 2006). When that face-to-face meeting does occur and the reality does not match the impression formed from the online interaction, it is usually the result of the participants “filling in the blanks” incorrectly (Goffman’s concept of idealization), not because they were mislead or lied to (Albright, 2001). People who meet and interact in a text-based environment are often wildly mistaken when they imagine one another’s off-line appearances (Jacobson, 1999), partially because of the impact of cultural stereotypes and personal schema on impression formation.

Ask Questions

Anticipated future face-to-face interaction also strongly impacts CMC participants and promotes the asking of more personal and intimate questions (Tidwell & Walther, 2002). Asking questions aids in self-disclosure and information gathering. If the available information is ambiguous, simply asking for clarification will ensure that accurate impressions are formed without jumping to incorrect conclusions. Asking questions can also reduce the possibility of the snowball effect or idealization occurring.

Precise Use of Language

The precise use of language is essential to managing impressions in CMC, to ensure that the expression equals the intention (Adkins & Brashers, 1995). Communications in a text-based environment can be extremely ambiguous, particularly because of the absence of nonverbal cues. For example, delivering the words “I hate you” with a smile, a wink, and an embrace make the actual intention of the message clear; paying attention to only the words used in the communication gives an inaccurate impression. Therefore, it is important to take advantage of the nature of CMC by taking the time to construct messages and craft them in the clearest and most comprehensible manner to avoid any misunderstanding.

Be Aware of Cues

Being aware of the verbal (text-based), linguistic, and paralinguistic cues present in CMC is the most important detail to pay attention to when working virtually. As discussed previously, the absence of nonverbal cues in CMC requires participants to use other content and linguistic strategies to gather
Impression Formation in Computer-Mediated Communication and Making a Good (Virtual) Impression

information that assists in impression formation. Virtual workers must guard against jumping to conclusions based on minimal cues as a guide.

Be aware of the subtle cues communicated by e-mail addresses; a virtual worker’s e-mail address should project professionalism. Think about the subtle meanings conveyed by linguistic style when interacting via CMC. Pay attention to details such as language intensity, lexical diversity, and spelling. Be aware of the impact chronemics can have upon impression formation. Paralinguistic cues can also communicate information that is used in impression formation, but only if the virtual workers involved in the interaction share the same framework for assigning meaning to the indicator.

Impression development is significantly dependent upon schema, context, and stereotypes, particularly if they are shared. Participants in a virtual interaction must have the same social context in which to work so communication is clear and understandable. For example, knowing the term “y’all” is associated with people from the southern part of the U.S. If that concept is not shared, the term is meaningless and provides no information that can contribute to impression formation. Impressions are based on the cues provided, in addition to cognitive models and conceptual categories as well as the context in which messages are viewed (Jacobson, 1999). Understanding how those verbal (text-based), linguistic, and paralinguistic cues can be interpreted in light of an individual’s cultural stereotypes and personal schema is key to managing impression formation when working virtually.

FUTURE TRENDS

The majority of research on impression formation in CMC has been conducted in text-based environments such as e-mail, bulletin boards and chat rooms. Newer technologies, however, allow for mediated interaction beyond mere text. Virtual workers can now communicate online using a variety of media and channels that permit the exchange of nonverbal cues.

Web cameras and videoconferencing allow real-time mediated face-to-face interaction where people can actually see each other as they engage in a dialogue. They can see the facial expressions, body posture, ethnicity and gender of someone as well as hear the inflections in a person’s voice. Digital photographs cannot provide real-time cues, but they still convey information such as the ethnicity, gender, and overall “look” (such as perceived attractiveness, hair color, manner of dress) of a person. Voice over Internet protocol (VoIP) allows people to talk to each other using a computer, where they can also discern accents, linguistic style and voice inflection.

Each of these “richer” (Daft & Lengel, 1986) communication technologies permits the transmission of nonverbal cues that reduce uncertainty. Common sense dictates that the use of these technologies will facilitate the formation of impressions in CMC. There is scant research into this area, but what little has been conducted has found otherwise.

Studies have shown that seeing a photograph of someone or interacting via videoconferencing does not facilitate and may actually hinder impression formation (Walther, Slovacek, & Tidwell, 2001; Fullwood, 2007), possibly because nonverbal cues may be distorted in a mediated environment, or those cues simply do not have the same impact in CMC versus face-to-face. Additionally, other nonverbal cues tend to be overemphasized in a mediated environment. A photograph may evoke a stereotypical impression, whereas the lack of a picture eliminates that possibility. Indeed, the loss of certain information can have positive consequences (Gilbert & Krull, 1988).

Even though richer communication technologies permit the transmission of nonverbal cues, they may not assist in impression formation. Therefore, virtual workers should still employ the strategies discussed previously to ensure
they make a good impression on their clients and colleagues when interacting using newer technologies.

**CONCLUSION**

Early research suggested a lack of nonverbal cues made CMC feel impersonal and would therefore hinder impression formation. Theories based on further research such as SIP and URT, however, indicate that people interacting via CMC can indeed form well-developed impressions (Sherman, 2001) by simply adapting and using other content and linguistic strategies to aid in impression formation. As Tidwell and Walther observe, “When people want to get to know one another, they overcome the limitations of the medium and do so” (2002, p. 341). Research has also shown that impressions formed in an online environment tend to be less detailed but more intense than those formed in a face-to-face interaction (Hancock & Dunham, 2001), but that impressions formed via CMC become more developed over time (Walther, Loh, & Granka, 2005). As uncertainty about a person is reduced, people form impressions and create mental models that help to understand the environment around them (Srull & Wyer, 1989).

People tend to believe they give the same impression to others regardless of whether the interaction is online or offline (Sherman, End, Kraan, Cole, Campbell, Klausner, & Birchmeier, 2001). As this chapter has demonstrated, that is clearly not the case. The absence of nonverbal cues that normally aid in impression formation drastically impacts how people form impressions in CMC. For virtual workers to gain social acceptance and develop and maintain viable relationships, they must employ specific strategies to carefully manage their online interactions:

- Disclose everything
- Be honest
- Ask questions
- Use precise language
- Be acutely aware of cues

“The impact of CMC on society is still evolving and our understanding of its social, psychological, political, and economic implications is far from complete” (Sherman, 2001, p. 69). As the workplace becomes increasingly virtual and the nature of business practices begins to change, the Internet is transforming how people interact and communicate. While the eventual societal impact of working virtually via CMC is still uncertain, one thing will always remain true: you never get a second chance to make a good first impression.

**REFERENCES**


Impression Formation in Computer-Mediated Communication and Making a Good (Virtual) Impression


Impression Formation in Computer-Mediated Communication and Making a Good (Virtual) Impression


KEY TERMS

**Computer-Mediated Communication**: Interaction that takes place through or is facilitated by mediating technologies, such as a computer.

**Exemplar**: A specific individual that a person has encountered, who is then taken as a representative of others who are thought to be in that same category.

**Impression Formation**: Creating an opinion or mental image of somebody or something.

**Impression Management**: The process by which individuals attempt to control others’ perceptions of them, motivated by the desire for social acceptance and relationship development and maintenance.

**Linguistic Communication**: Interacting and gleaning information using language, both verbal and textual.

**Nonverbal Communication**: Interacting and gleaning information not using words but by using cues taken from actions such as vocal patterns, linguistic markers, body posture, gestures, and eye contact.

**Paralinguistic Cues**: Typographical marks and other textual features that have no lexical meaning per se, but the meanings are dependent on group or individual contexts.

**Stereotype**: Where a society or culture recognizes characteristics of an individual or group of people as representing an entire category.

**Working Virtually**: Interacting and communicating to achieve set goals or tasks using information and communication technologies, not in a face-to-face manner.
Chapter 5.12
Group Decision Making in Computer-Mediated Communication as Networked Communication: Understanding the Technology and Implications

Bolanle A. Olaniran
Texas Tech University, USA

ABSTRACT

Networked communication is proliferating our world. The fact that global information communication technologies (ICTs) are becoming increasingly available is facilitating human computer interaction, which permeates the use of computer-mediated communication (CMC) in various organizations, groups, and interpersonal contexts. As a result, the issue facing today’s organizations is not whether to use global information technologies (GITs) in networked communication, but rather how to use them for effective functioning and as efficient coordination tool; especially how to incorporate GITs into the decision-making process. Consequently, this chapter examines the issues in designing CMC into group interactions and decision-making processes.

INTRODUCTION

The shift toward globalization in organizations necessitates increase reliance on information communication technologies (ICTs). However, the nature of ICTs required is not the stand alone, but rather networked ICTs. The growth in networked ICTs such as e-mail, computer conferencing, desktop conferencing, and videoconferencing emanates from the need for speed in decision
Group Decision Making in Computer-Mediated Communication as Networked Communication

Making and the general ability to coordinate activities of geographically dispersed workers, or to facilitate learning beyond certain geographical boundaries. The networked communication process and the accompanying decision-making processes are made possible by the Internet.

**BACKGROUND**

Many organizations and business executives continue to realize the role of employee participation in the decision-making process and organizational productivity for survival. Thus, organizations continue to develop ways to actively engage their members in teams either in co-located or distributed virtual groups. Networked global information technologies (GITs) especially benefit geographically dispersed groups to engage in decision-making process virtually. However, the literature addressing GITs and computer-mediated communication (CMC) focus on comparative features of the technologies over face-to-face (FtF) or traditional communication media instead of looking at the decision-making process occurring over these media as instances of networked communication. There is a potential benefit in examining decision making as networked communication. This approach allows one to ask questions about what technology to use, or why use a particular technology, and when to use a given communication technology? Therefore, the goal of this chapter is to offer insights about the decision-making process in ICTs by looking at some key factors that must be addressed, while offering some recommendations regarding the effective decision-making process when using GITs. First, however, the next section identifies and discusses factors in networked communication decision making as a way to address issues and problems in the networked communication.

**ISSUES, CONTROVERSIES, AND PROBLEMS**

Although the group decision-making process is made possible in networked communication, the tendency to succeed or fail hinges on attention to some key issues. For instance, there is a greater need to adapt communication technology and communicative behaviors to accommodate the decision-making process across different geographical boundaries and social contexts (Olaniran, 2004). The decision-making process where group members are located across different geographical boundaries and time zones necessitates the need to attend to culture. Attending to different cultural needs is not only necessary, but also is increasingly essential for an effective decision-making process to occur. It has been suggested, that success in organizations at large has less to do with market forces than it does with the cultures and the values existing within cultures (Alas & Vadi, 2003; Cameron & Quinn, 1999). There are specific instances of the culture factor in networked communication and the decision-making process. A case study identified that the East Asian culture and its social structure showed cultural differences through the suppression of e-mail use when subordinates are interacting with superiors. More specifically, subordinates are reported to refrain from the use of e-mail with superiors. This is not an isolated incident, as a similar difference was found with the Dutch who showed greater preferences for a more structured decision-making process than their American counterparts (Kiser, 1999; Gezo, Oliverson, & Zick, 2000).
Attention to the issue of culture is important in networked communication for two reasons. One, culture preferences may affect the choice of GIT or communication medium for interaction. Two, the need to realize that technology (i.e., networked ICTs) may not necessarily overcome some deeply held traditional beliefs that underlie how individuals communicate.

Culture is also important in networked communication because it interferes with interactions of groups in organizational decision-making processes. For instance, people in co-located groups have greater access to multiple communication media and, thus, have the benefit of using multiple channels, which in turn permits a broader range of messages, cues, and feedback immediacy (Olaniran, 2004). This is not the case in networked group decision-making processes. Even when individuals have access to other communication media, their usage may be impossible due to geographical distance, and thus, clarification that facilitates message comprehension may be hindered with delayed feedback (Olaniran, 1995). For example, Armstrong and Cole (2002) found that remote sites in virtual teams often fell off radar screens and usually were ignored during telephone and videoconferences. In addition, in videoconferencing, unless the system has cascading capacity, the most visible locations participants are able to see are restricted to two (i.e., picture in picture). Furthermore, time zone differences also create some sense of psychological distance along with the problem of finding a time that works for all group members.

Similarly, the temporary nature and short history of many geographically dispersed teams has become the center of attention in development and sustenance of the relationship among members in CMC decision-making groups (Poole & DeSanctis, 1992). For example, consensus building is one such area in decision making. It is argued, that when mediated group members in dispersed groups have limited time, they often fail to seek adequate social and contextual information to support their attributions (Crampton, 2002; Olaniran, 1994; Walther, 2002). This may result in dire consequences for participants in maintaining relationships or developing cohesion in dispersed CMC groups. Without a shared frame of reference that is developed through frequent and previous interactions, participants encounter problems in decision-making groups. Accuracy in interpreting group members’ messages becomes difficult (Olaniran, 2004). The problem is expected to occur in unitary organizations operating under the same cultural norms, while it is expected to be heightened or exacerbated in distributed networks and in organizations that are geographically dispersed (Kelly & Jones, 2001; Olaniran, 2004).

It is important to examine some of the reasons why decision making over GITs is attractive to organizations that are committed to it and why others are clamoring to explore it. The next section looks at the key benefits of computer-mediated communication and other GITs.

**Some Key Benefits of CMC**

**Freedom to Participate**

CMC has been touted as the means to bring about freedom of participation and increased interaction in groups (McComb, 1994; Olaniran, Stalcup, & Jensen, 2000). CMC discussions offer greater freedom and opportunities to participate than face-to-face (FtF) meetings (Phillips & Santoro, 1989; Olaniran, 2004). In general, the greater participation in CMC (e.g., e-mail or computer conferencing) is attributed to the concurrency of the communication medium. Concurrency is the ability of a communication medium to allow different individuals to simultaneously contribute to the decision process in network communication (Valacich, Paranka, George, & Nunamaker, 2004).
In other words, group participants do not have to wait until a group member completes a thought or yields the floor before they can add their comments. Thus, participants can input their comments at the same time without interfering with other members’ messages especially in synchronous interaction. In an asynchronous interaction, the technology still allows members to participate in the decision process at their convenience by allowing individuals to work around different schedules and time zone differences. This is not the case with FtF. Participation in the decision process through FtF occurs with the requirement that all group members be in attendance to contribute their ideas. Also, FtF interaction requires members to take turns in order to have meaningful interactions.

Equalized/Democratic Participation

Another benefit of the freedom to participate or increased participation enhanced by CMC is the possible equalized or democratic participation. Some scholars have argued that CMC increases the democratic decision-making process (Dubrovsky, Kiesler, & Sethna, 1991; Phillips & Santoro, 1989). This is the area where characteristics and features (tools) designed into different GITs come into play. For instance, certain decision-making tools, such as voting and analyses tools, are offered in some decision support systems that allow groups to reach consensus on decisions (Poole & DeSanctis, 1992). In addition, videoconferencing, through the ability to see co-communicators improves participation and interpersonal interaction as a whole (Badenhorst & Axmann, 2002). However, one must not forget the fact that it is the choices that people make about GIT features regarding their appropriation or use that will determine the final outcomes such as facilitating democratic decision-making process or equalized participation. For example, Poole and DeSanctis (1992) found that the choices people make in using CMC systems often counter the intent of the system designer. In other words, CMC features are designed to accomplish the spirit, which is a particular purpose, but those features, are applied in unintended ways. More specifically, groups appropriate systems from, not only the structural features of the technology but also from other relevant institutions, such as the environment and from group characteristics (Poole & DeSanctis, 1992). Furthermore, user knowledge affects the way tasks are approached in CMC. For example, it has been shown that inexperienced users are more interested in task accomplishment, while more experienced users are interested in efficient application of ICTs (Trumbly, Arnett, & Johnson, 1994). The former stressed learning the minimum requirements for task performance, while the latter put greater emphasis on comprehension of technological features. Consequently, the resulting outcomes in one group over another may vary to a certain degree. The next section will address conditions for which to use ICTs and how to use them successfully.

SOLUTIONS AND RECOMMENDATIONS

Given some of the possible contextual variations in decision making and communication outcomes in networked communication, it is useful to address conditions when certain GITs can be used to foster effectiveness and efficiency. First, effectiveness, in a general sense refers to the idea of process or usage leading to accomplishment of group goals. Second, effectiveness, in a specific sense, refers to the opportunity that GIT offers to improve what is possible under the traditional method (e.g., FtF). Efficiency, on the other hand, focuses on factors such as the speed or cost at which access to information is attainable. Therefore, deciding what technology to use and when to use it is possible by focusing on these two factors.
In light of contextual variations in technology use, there are certain conditions under which equality may result from CMC interactions. For instance, it should be noted that anonymity usually neutralizes any status effect and, thus, allows participants to contribute freely to discussion. Individual identity is protected, and participants do not have to worry about how their messages will be received (i.e., evaluative apprehension) by other participants (Olaniran, 1994). However, the challenge with anonymity is that even when organizations attempt to safeguard members’ anonymity in the decision-making process within CMC, users still may be apprehensive about their contributions. Furthermore, while anonymity is available under experimental conditions, most organizational use of CMC is devoid of anonymity. Yet, a certain degree of equal participation still may be achieved in such settings. For instance, when the communication setting surrounding the problem solving and decision making is informal (e.g., when participants, organizational, and group members are on a first-name basis), interactions that overlook status are possible, enhancing freedom to voice opinions and increase the level of attention paid to opinions and suggestions. In other words, organizations should pay attention to developing open communication climates that would carry over into interaction over technology communication networks.

The need to facilitate an open communication climate and increase participation in group interactions also involves attention to the social norms that guide communication, the structure, and rules, or organizational cultures that guide acceptable communicative behaviors. This argument is in line with the finding that when GIT is used in conditions where group norms are salient, members tend to exhibit behaviors that conform to accepted group norms (Lea & Spears, 1992). Thus, it is necessary to note how social contexts like rules and norms may take precedence over medium effects.

Furthermore, unlike dyadic relationships, group interaction focuses attention away from independent users (i.e., sender and receiver of messages) toward interdependent relationships. Therefore, if a group member selectively discards someone’s message in CMC, others can re-introduce the message during the discussion. This is something that might be too threatening in FtF, especially when the decision process involves interaction between high- and low-status individuals. In essence, CMC has the capacity to mask cues attributable to power and status differences, and therefore, increases the freedom to exchange information (Siegel, Dubrovsky, Kiesler, & McGuire, 1986).

Consequently, CMC holds an advantage over traditional FtF interaction in reducing negative impacts of selective message processing or redistributing power among group members. This argument is consistent with the network paradigm presented by Rice (1990), which emphasizes principles of convergence, where attention is redirected from individuals as independent senders and receivers to one where an individual is an actor in a network characterized by interdependent relationships in social settings. To this end, Kelly and Jones (2001) contend that managers and organizational decision makers pay as much attention to the underlying development and the social communication structure as they do to GIT infrastructure.

The role that social cues play in decision making and communication interaction needs to be addressed. The degree with which GITs allow social cues in interaction differs depending on the type. A general criticism of text-based GIT is that it is lacking in nonverbal cues. For instance, research on relational communication argues that nonverbal codes are relationally warm and fulfilling while arguing that verbal codes are best at satisfying the content or task function (Walther, 2002). Notwithstanding, the lack of nonverbal cues in certain technologies, GITs, have been shown to
encourage participation by masking status effect that is likely in FtF discussion. Furthermore, the lack of nonverbal cues in CMC does not equate an absence of social cues from the communication process. For example, different forms of social cues have been found in CMC: Paralinguistic cues that involve textual manipulation to add context and considerable socio-emotional content to messages (Lea & Spears, 1992; Rice & Love, 1987). Others consist of lexical surrogates and icons used to produce codes or symbols that convey expressions similar to nonverbal facial expressions such as the smiley. More recent examples include sophisticated emotions with audio capability like the ones employed in the Yahoo instant messengers and other computer conferencing. Furthermore, paralinguistic codes such as capitalization, parenthetical notes, and exclamation marks are used to express emotion and meaning (Lea & Spears, 1992).

There also is empirical evidence that confirms text-based CMC groups demonstrate socio-emotional tone from the moment of initial interaction (Walther, 1997). At the same time, advances in computing capability, graphical emoticons are improving how social information is conveyed in CMC.

Making Networked Communication Work

The choice and selection of GITs must be based on the overarching goal of the group. The two-stage model of the group decision-making process may be used to offer some guide (Olaniran, 1994, 2004). The two-stage model consists of the idea generation and idea evaluation. The idea generation stage involves groups attempt to analyze problem, determine goals, and generate possible solutions for a given task. The idea evaluation, on the other hand, focuses on developing criteria for evaluating possible solutions, and discussions of implications of possible decisions and the eventual consensus on selection of solution to the task.

The need for member creativity is particularly important during the idea generation stage of the decision-making process. In order to accommodate creativity, interaction procedures that maximize participants’ participation as alluded to earlier are important (Olaniran, 1994, 1995). Consequently, GITs that nurture individual creativity while reducing evaluation apprehension would be the most productive. It would appear that any GIT would do, provided the organizational culture and interaction norms stress informality. However, when, this is not the case, organizations and groups should strive to use text-based CMC with special attention to protect participants’ anonymity. In this case, networked communication that utilizes central bulletin board or discussion list where ideas are less likely to be traced back to individual contributors. This way, group members are able to participate without fear of reprisals. Furthermore, they will not be participating as in FtF where nonverbal cues can negatively influence group member’s level of participation. Specifically, Olaniran (1994) found that anonymous CMC groups generated more than three times as many ideas as FtF groups and almost two times as many ideas than other CMC groups when group members do not have to worry about their identity being revealed.

The decision to select either synchronous or asynchronous CMC should be determined by the amount of lead time a group has to complete the task, as it would seem to have no effect on the idea generation stage. However, the need to offer anonymous participation holds certain implications for the degree to which e-mail, intra-organizational instant messengers, teleconferencing and videoconferencing can be used in the idea generation stage of the decision model. These communication media clearly reveal the user’s identity in networked communication. The way around this problem, it seems, is to have in place organizational and group climate that stresses informal structure.
The idea evaluation stage, contrary to the idea generation stage, focuses on the role of critical thinking in the process of enhancing decision quality. Thus, the need for group members to freely discuss the pros and cons of each possible solution is crucial. At the same time, the need to develop social interaction and bond (i.e., cohesion) with other group members is important and best nurtured during the idea evaluation stage.

Therefore, communication media that can facilitate critical analyses of ideas while helping in addressing relational and social needs of group members are needed during idea evaluation stage. The development of a lasting relationship while reaching high quality decision making is best accomplished in a group when the social structure such as breaking down cultural barriers, removal of perceived distance, and utilization of GIT media that can appeal to different senses are in place. For example, Waldir Arevoedo (a research analyst at Gartner Inc.) argues that the more CMC interactions involve multiple media that can support cues such as voice, text, and audio, the more the barriers of time, distance, and culture can be overcome (Solomon, 2001). Therefore, ICTs such as videoconferencing and audio teleconferencing offer the most cues for accomplishing relational goals in the decision-making process and networked communication.

An alternative, is to provide multiple communication media or ICTs that allows participants to take control of how they choose to participate in the decision-making process. Thus, when group members need social relations they can select the medium that best accomplishes the desired goal. However, at times, these communication media are not accessible (Olaniran, 2004). When access to multiple communication media, or those providing nonverbal cues, are not the case, attempts must be made to manipulate text-based CMC to meet the need of group members. One method offered for developing social relations is to allow members to work in ongoing groups where group members are able to continue to negotiate and renegotiate their relationships over different tasks (Olaniran, 2004). This suggestion may also help foster critical thinking needed in idea evaluation of the decision making by allowing members to refrain from free riding or lurking activity.

As in the idea generation stage, the decision to use synchronous or asynchronous networked communication should depend on the amount of time a group has. Notwithstanding, synchronous networked communication fosters increased immediacy and feedback, such that clarification of ideas and answers to group member concerns are instantaneous. Similarly, synchronous networked communication allows the decision-making process to flow more smoothly than in asynchronous interactions. On the other hand, idea evaluation in asynchronous networked communication offers group members more time to reflect over ideas before composing a response. 

**FUTURE TRENDS**

One of the major hindrances to adoption and usage of global information systems in organizations’ decision-making processes around the globe, involves the propensity to transfer and use communication technology. The propensity to transfer and use communication technology involves the degree to which global and multinational corporations are willing to allow their technologies to become accessible to their joint venture partners in another country or world region. The problem with propensity to transfer and use technologies has even far greater consequences for the less economically developed countries (LEDCs) where infrastructure and cultural values are different. In essence, the propensity to use global information systems begins with access to technology and the willingness to use it. For the most part, the lack of adequate access to certain global information systems, creates a digital divide between individu-
als from economically developed and those from the less economically developed countries.

Bozeman (2000) argues that technology transfer is based on cost and benefits, and usually the transfer exists only when benefits outweigh the cost. This is significant in the race to use GIT and in bridging the digital divide between economically developed countries (EDCs) and LEDCs. Major information technology advances occur in developed countries, to the detriment of LEDCs, such that affordability becomes a critical issue. Along the same line of reasoning, Ya’u (2004) argues that LEDCs are primarily consumers of global information systems and technologies. As consumers of GITs, LEDCs face knowledge deficiency that also could affect creative use of technologies. Furthermore, there is different motivation regarding the propensity to transfer technologies. While the United Nation’s Commission on Science and Technology for Development wants to use ICT to bridge the digital divide by creating human development initiatives in LEDCs, World Trade Organizations (WTO) and traditional organizations are in it for profit motives in what has been described as another attempt to re-colonize LEDCs in the new global order (Ya’u, 2004). Consequently, the motivation to transfer technologies is not aimed at bridging the digital divide, but, instead, it is driven by conglomerates’ needs to access other markets in response to their home markets, which are increasingly inadequate for sustaining required corporate growth due to saturation of the developed countries’ markets (Ya’u, 2004).

Furthermore, Ishemo (2004) claims that new information and communication technologies are invented in the framework of capitalism development. Consequently, foreign direct investment and critical technology transfer goes to profitable markets rather than markets where there is the need to promote universal access by subsidizing access to global information systems. These strategies perpetuate the divide rather than actually bridging it. For instance, it was found that international firm, large host country, and recently established affiliates correlate highly with technology transfer. Grosse (1996) stresses the profit motive behind technology transfer when indicating that technology transfer is greater when the cost of carrying out the transfer is low and when the resulting benefits of technology transfer were higher. Similarly, host countries’ local government regulations that govern partnerships and joint venture programs and political instability in LEDCs creates additional risks for technology transfer, which some companies are not willing take. Organizations fear that, unless the revenue potential is substantial in a given market, their investment in the country may be short-lived.

At the same time, when technology is available, the propensity or motivation to use it is an internal one. There has to be a willingness or motivation on the part of potential users to want to try or use them (Storck & Hill, 2000). The motivation to use a particular technology is linked to the cultural norms in place. Some cultures resist technology or adopt technology only as long as it does not conflict with their cultural norms (e.g., Heaton, 2001). The example of Cuba’s application of the Internet illustrates this point. Also, older generations have the tendency to resist new technology because of complacency with old ways of doing things (Wheeler, 2001). At a specific level of culture is the issue of language, where over 87% of Internet content is in English (Wilborn, 1999; Van Dam & Rogers, 2003). Non-English-speaking individuals may feel that technology has nothing to offer them. It stands to reason that people in poor nations or rural communities would be motivated to learn and use technology when it helps them access information tailored to their specific needs (see Sassi, 2005).

Certainly, realization of benefits offered by any global information system, depends on the ability of the users to explore and experience proposed benefits on their own. Forced usage would result
in resistance and rejection of the technology as a whole and, more specifically, it would create failure to achieve successful incorporation into the decision-making process.

**Implications and Recommendations**

The driving factor in decision-making processes (networked or otherwise) is the need to attain a level of trust that is essential for group members to reach consensus and effectively make a decision. Schein (1992) argues that decision makers must value and develop relationships that would facilitate joint problem solving and solution implementation. The goal in decision-making groups is to foster adaptability, creativity, and flexibility in group interactions, which have direct impact on organization learning. In connection with learning, it has been suggested that a focused task-oriented culture is crucial to cope with the external environment, while a relational-oriented culture is crucial for accomplishing internal integration (Alas & Vadi, 2003). However, developing organizational learning also demands the ability to work together as teams (Alas & Vadi, 2003; Senge, 1997). Thus, a sense of common identity is developed when each group member serves as reference point for one another. Group participants’ dependence on one another for socio-emotional, as well as task support, enhances relational development that indirectly influences commitment and overcomes barriers to learning.

The information provided in this chapter suggests that the issue facing modern organizations is not whether to use ICT or CMC in the decision-making process. Rather, the challenge is how to adapt communication technology in group collaboration and the decision-making process effectively and in a way that fosters task and relational goals respectively. In order to accomplish such a complex task, attention must be paid to certain environmental factors. First, organizations embarking on implementing networked communication in its group decision-making process would need to realize that there are social obstacles that technology or CMC is unable to overcome. For instance, technology can be used to facilitate relationships within organizations provided individuals seek out such relations. Otherwise, technology in and of itself would not create trust, creativity, and adaptability needed for success in group decision-making processes. Second, the decision to use GIT in a group process need not be based on the idea of “everyone is using it,” rather, it must be evaluated based on the fact that its deployment will help accomplish goals that are central to organizations’ core practices. Third, the social context in which GITs are being deployed is just as important as the task for which they are deployed to accomplish. Differences in outcomes from research in GIT and group decision processes are significantly influenced by how individuals adapt technology, which is often different than GIT designers’ intent.

Organizations’ need to recognize that GIT and other groupware provides a part of the overall communication environment (Kelly & Jones, 2001). It is important to understand that GIT and other decision support are available to supplement face-to-face and other traditional media and not replace them. It is also recommended that software designers need to help organizations overcome some of the challenges in GIT. One is the issue of bandwidth, which prevents against implementing more sophisticated GIT applications, such as those allowing audio and video channel cues, the likes of computer videoconferencing and IPTV. Along the same line, designers need to create third-party software that offers multiple cues that are easy to deploy in the context of group decision making, and are easy to access across the globe.

Networked communication media designers are encouraged to keep in mind criteria involved in the decision-making process by paying attention to the stages, goals, and social needs of organizations and users. One way to bring
this about is to create GIT architecture in a way that adheres to the “analytic hierarchy process” (AHP) of decision making at the minimum. The AHP involves decision-making communication protocols that conform to real-world problems (Ghinea, Magoulas, & Siamitros, 2005; Saaty, 1980). The AHP approach is similar to the two-stage model of decision making by emphasizing that a networked communication media allows users to be able to generate sets of alternatives, prioritize those alternatives, address issues of resource allocations, and predict possible outcomes while resolving conflicts. The capacity to address subjective criteria in the decision-making process also is considered as paramount in any systems design. The AHP should be taken as minimum criteria upon which other features can be incorporated into GITs.

Along the same line, Olaniran (2004) recommends a combined media in virtual or networked groups. He suggests that the combination of both FtF and CMC media in decision making offers an approach that gradually introduces group members to GITs and also creates an approach where GIT users do not feel like they are completely giving up face-to-face interaction for electronically mediated communication. Finally, the need to address the needs of GIT users is important to the decision-making process as much as the decision outcomes from the use of these technologies.

**FUTURE RESEARCH DIRECTION**

Future research on global information technology (GIT) must begin to explore communication technologies not as “either or” when compared with traditional FtF medium. Rather the unique characteristics of each communication technologies should be assessed and then effort should be made to identify how these characteristics can complement established and effective communication interaction and decision-making processes. Furthermore, it is essential that new research focuses on cultural implications of communication technologies. For instance, questions such as: How different communication technologies reinforce or disrupt clearly established cultural patterns of communication across the globe? What effects do dimensions of cultural variability have on organizations’ struggles with communication media selection and usage? These questions are essential in the globalization age where product and service standardization along with market differentiation (i.e., globalization) is the norm.

Along a similar trail of thought is the need for researchers to explore how best to use new and modern CITs to reinforce existing organization culture. It has been found that CITs do not excel when used to change organizational cultures, but instead, they help promote existing cultural norms (e.g., Canessa & Riolo, 2006).

Future research also needs to focus on the use of certain CITs (e.g., videoconferencing and synchronous CMCs). While these CITs are purported to add immediacy and in a videoconferencing case, geographical distances in terms of different international datelines may render them useless. Similarly, it would help to examine the impacts of these synchronous CITs on actual member participation in the decision-making process and as communication media of choice by organizational members. The issue of anonymity or perception of anonymity makes using asynchronous CITs attractive to people for participation in the decision-making process. Thus, there is the need for future research to explore implications of synchronous interactions over CITs when anonymity is removed especially within different cultures.

Taken together, addressing these areas would add significant information to the body of literature in GIT. At the same time, it will help researchers on theory development. As for organizations and their members, the information would be invaluable in identifying what works and how to effectively incorporate CITs into the decision-making process.
CONCLUSION

Finally it is essential that as executives in their respective organizations attempt to deploy or incorporate GITs into their decision-making processes, it would help for them to pay attention to how users in different cultures react to these technologies. Additional impetus could come from researchers who would be able to evaluate meaningfully the role of different cultures in GIT deployment. For instance, while the idea of democratic participation and freedom to contribute in the decision-making process is welcome in the Western culture; would individuals in power distant culture (where status is cherished as the cultural norm) welcome this approach? In other words, there are several cultural factors, including the exercise of power, in terms of who has it, and how they guard it, that may hinder effective implementation of global information technologies in decision-making processes that are beyond the scope of this chapter. However, this chapter offers a potentially successful platform for executives to adapt and modify as situations demand.

REFERENCES


Crampton, C. D. (2002). Attribution in distributed work groups. In P. Hinds & S. Kiesler (Eds.), Distributed work (pp. 191-212).


Heaton, L. (2001). Preserving communication context: Virtual workspace and interpersonal space in Japanese CSCW. In C. Ess & F. Sudweeks (Eds.), Cultural attitudes towards communication and technology (pp. 163-186). Australia: University of Sydney.

Group Decision Making in Computer-Mediated Communication as Networked Communication


**ADDITIONAL READING**


This work was previously published in the Handbook of Research on Global Information Technology Management in the Digital Economy, edited by M. Raisinghani, pp. 243-256, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
The purpose of this study was to answer the question: Is the level of achievement of working adult students in an online class directly related to the method of computer-mediated communication used by faculty? The study examined the relationship between the methods of computer-mediated communication utilized, the independent variable; and student outcomes, the dependent variable, among working adult students in online courses. Through an examination of course communication records and student final exam grades, the researcher developed course-based measures of the methods of computer-mediated communication and student outcomes. These measures were used to conduct statistically appropriate tests to determine if there was a statistically significant difference in the student final exam scores between classes that used only basic methods of computer-mediated communication as opposed to courses that employed both basic and advanced methods.

With businesses continuing to increase their expenditures for employee education, there is progressively more focus on maximizing employees’ educational outcomes to effectively leverage corporations’ investment. Dunn (2000) predicts that large corporations will develop their own approval systems for higher education programs similar to the current regional accreditation process.

Bradburn and Zimbler (2002) make the point that the 1999 National Study of Postsecondary Faculty lacked detailed questions about modes of technology, training and instructional practices in individual distance education courses and affirm that further studies in faculty participation in distance education are needed. Northrup (2002) confirms this opinion by stating that future studies should consider variables that may affect the individual learner, the learning environment and instructional strategies that may be most appropriate for specific learning outcomes. Quilter and Chester (2001) emphasize that few formal
Effects of Computer-Mediated Communication

research studies have been conducted to examine the relationships between online communication technologies and teaching and learning, and reaffirm that research employing empirical documentation of communication technologies is lacking.

The method of computer-mediated communication is important since different technology-supported methods of interaction (e.g., two-way interactive TV, text-based chat, e-mail) have different characteristics in regard to immediacy of feedback, student-student and student-faculty interaction, realism and student user control (Smith & Dillon, 1999). Kearsley (1995) reinforces the belief that the method of computer-mediated communication used is critical to the learning process, as it affects the provision of feedback to the student. Distance learning environments that support synchronous communication can provide immediate feedback to the learners and rapid interaction between learners, a feature that may motivate some learners. In contrast, distance learning environments that support asynchronous communication can provide the student with more control over where and when communication occurs as well as more time to reflect on and respond to course content and communications (Moore & Kearsley, 1995). The modality of communication employed in the course is a significant factor in determining the nature of the learning community that is formed in the course.

Duffy and Kirkley (2004) state that most higher education research relies on survey data; that is, class ratings and specialized surveys like the National Survey of Student Engagement (Kuh, 2001), to infer, based on student report, that learning has occurred. It is this researcher’s opinion that one of the main reasons for a lack of prior studies using quantitative data collected directly from course records has been the difficulty inherent in obtaining such data in an online, distributed learning environment.

THE RESEARCH PROJECT

The research project consisted of a combination of three individual studies covering a total of 116 courses and approximately 1,700 students that, taken together, provide a detailed analysis of the question. The first study of 16 courses provided a proof of concept to test the instruments and measures developed and to identify any significant issues. The main research effort involved a large sample of 80 courses that was built upon the results of the preliminary study. The third and final study of 20 courses was a validation study to further evaluate the repeatability of the results and assure the reliability of the instruments and measures.

For purposes of this project, computer-mediated communication was defined as: computer-mediated communication between instructor and student or between students, which discussed some aspect of course content, assignments or student progress in an online course. The study examined outcomes of courses that were supported by both synchronous (simultaneous) and asynchronous (delayed) methods of computer-mediated communication.

Data collected from course management software administration statistics included: (a) methods of computer-mediated communication employed and (b) student final exam grades. Data analysis was performed using appropriate statistical techniques, including the use of an index to represent study variables. The use of an index allowed for statistical analysis and comparison of the aggregated levels of activity. The following index was created: per course combined student final exam grade index (STUFIN).

STUFIN was defined as the arithmetic mean of the student raw percentage scores achieved on the final exam for a specific course. It is the average score received by students on the final exam on a scale of 0-100%, with 100% being a perfect score.
Software features provided by the course management software and used by new online faculty were divided into two categories; basic methods of computer-mediated communication and advanced methods of computer-mediated communication. The methods were grouped as either basic or advanced based on the complexity of the method, multi-media aspects of the method and the method of utilization of the specific method in delivering the online course (Table 1). Courses designated as utilizing basic methods of computer-mediated communication employed only those items listed as basic methods in Table 1. However, courses designated as utilizing advanced methods of computer-mediated communication employed both basic and at least some advanced methods of computer-mediated communication.

**Table 1. Basic vs. advanced methods of computer mediated communication**

<table>
<thead>
<tr>
<th>Basic methods</th>
<th>Advanced methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threaded discussions</td>
<td>Interactive chat rooms</td>
</tr>
<tr>
<td>Course Gradebook, including detailed feedback to students and opportunity for student/faculty interaction and postings related to a grade</td>
<td>Online Whiteboard</td>
</tr>
<tr>
<td>Course introductory announcements</td>
<td>Group teleconferences</td>
</tr>
<tr>
<td>Animated graphics (GIF)</td>
<td>Customized course calendar</td>
</tr>
<tr>
<td>Class announcements</td>
<td>Online interactive simulations or project labs</td>
</tr>
<tr>
<td>Lecture notes/PowerPoint presentations</td>
<td>Video screen capture/audio or video presentation download</td>
</tr>
<tr>
<td>Webliography</td>
<td>Streaming audio/video embedded in body of course material</td>
</tr>
<tr>
<td>Document sharing</td>
<td></td>
</tr>
</tbody>
</table>
The final exam process was designed so that final exams were administered consistently and included comprehensive coverage of the entire course. The consistent nature of the final exam structure and process throughout the university and the wide range of disciplines and number of classes included in the study served to increase internal validity.

RESULTS

In each of the studies comprising this project there was a statistically significant difference in the average final exam score achieved for students in courses utilizing both basic and advanced methods of computer-mediated communication vs. the average final exam score achieved for students in courses utilizing only basic methods of computer-mediated communication. The statistical analysis was done in two stages: first using a student’s t-test for comparison of the group means of the STUFIN index, and then using ANOVA to compare the group means. In each of the studies the results of both types of analysis were significant at the .01 level, indicating that there was a statistically significant difference. The advanced methods courses always returned higher average final exam scores. The average values for the STUFIN index are shown in Table 2. The t scores, ANOVA F scores and related significance levels for the comparison of the group mean values for STUFIN in each study are shown in Table 3.

Smith and Dillon (1999) refer to the media/method confound, a concept stating that the technology alone does not cause the effect; rather, it is the combination of the technology and the way the technology is employed that impacts student outcomes. The act of utilizing advanced features in a course indicated that the faculty member was attempting to actively employ technology to provide a better learning experience for the students. It should be noted that the mere presence of advanced software features was not sufficient to affect a change in student outcomes. Those courses that employed only basic methods of communication had all of the advanced features and functionality available, though were not utilized. The presence of more sophisticated technology does not translate to more effective results unless that technology is employed in an effective manner.

The results indicate that as a group, students achieved higher final exam scores in courses in which the faculty member made an effort to employ the advanced technology in the course.

Table 2. STUFIN average values

<table>
<thead>
<tr>
<th></th>
<th>Basic and Advanced Methods of Computer Mediated Communication</th>
<th>Basic Methods of Computer Mediated Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preliminary Study</td>
<td>82.96</td>
<td>73.01</td>
</tr>
<tr>
<td>Main Study</td>
<td>82.57</td>
<td>70.87</td>
</tr>
<tr>
<td>Validation Study</td>
<td>83.90</td>
<td>72.19</td>
</tr>
</tbody>
</table>
The effectiveness and manner in which a given faculty member employed the more advanced software features may have contributed to the amount of improvement in student outcomes realized in a given course.

**LIMITATIONS OF THE STUDY**

The primary limitations of the study included the experience level of the faculty member with technology and specifically with online educational technology, a factor that may have potentially impacted the internal validity of the study results. This factor was mitigated by limiting the faculty group to new online faculty. All new online faculty participating in the study received (a) identical on-site training in the use of the software toolset, (b) identical Internet-based training in the use of the software toolset, and (c) individual ongoing mentoring throughout the initial course.

The experience level of the student with technology and specifically with online educational technology also may have impacted the internal validity of the results. Students with significant experience and comfort using online educational tools may participate in a course to a greater degree than less experienced students. Incorporating a wide mixture of courses ranging from beginning to advanced levels served to mitigate this effect.

There may be a number of indirect relationships, supported by multiple factors, impacting outcomes in online classes, including (a) student perceptions and attitudes or (b) institutional requirements, such as the required minimum level of participation in class. The analysis of these indirect relationships was outside the scope of the proposed study. The relative effectiveness of different advanced methods of computer-mediated communication or of an individual instructor’s ability to employ the software was also outside the scope of the study.

**CONCLUSION AND FUTURE TRENDS**

As course management systems become increasingly sophisticated, the associated acquisition, licensing and maintenance costs will also tend to increase. Both of these trends—the increasing sophistication of the course management software and the increasing cost of that software—appear to be ongoing.
The study results suggest that the investment in these software systems is, in spite of the cost, a proper decision that can be of value for both the university and the student. The difference in student final exam results is a clear indicator that this ongoing evolution is a positive factor that can contribute to the success of online education programs.

Garrison and Shale (1990) state that the quality and effectiveness of education at a distance is directly attributable to the kind of communication between teacher and student. The value that advanced software features provide is in their ability to offer faculty more options and creative alternatives in communicating with the student and delivering the course material.

Future research regarding reasons for the preference of a specific method of software-facilitated communication or a specific software feature may assist in developing future enhancements to functionality in course management software. The inclusion of all interested parties in the ongoing software development process should assure that the needs of business, students and educators are all being addressed.

This study was conducted with working adult students enrolled in an accredited university. A research area that merits further study is the relationship between methods of software facilitated communication and student outcomes in online programs with different demographic profiles. Additional research might also be conducted into the relationship between the method of computer-mediated communication employed and student retention in online programs.

NOTE

An earlier version of this paper material has been published in the Journal of Educators Online, Volume 1, Number 1, Summer 2004.

REFERENCES


Northrup, P.T. (2002). Online learners’ preferences for interaction. The Quarterly Review of Distance
Effects of Computer-Mediated Communication


KEY TERMS

Advanced Methods of Software Facilitated Communication: For purposes of this study, advanced methods of software facilitated communication were defined as including features such as interactive chat rooms, whiteboards (live interactive chats), group teleconferences, customized course calendars, interactive simulations and virtual project labs.

Asynchronous Communication: A delayed time communication, typically in text format, between the learner and instructor or among learners.

Basic Methods of Computer-Mediated Communication: For purposes of this study, basic methods of computer mediated communication included such tools as threaded discussions, grade-books, class announcements and lecture notes.

Computer-Mediated Communication: Communication between instructor and student or between students, which discusses some aspect of course content, assignment or student progress in an online course, utilizing the online computing environment for the communication.

Course Management System: A software-based system for managing the development and delivery of online courses and programs and managing student progress over the Internet.

Final Exam Grade: For purposes of this study, final exam grade is defined as a student’s percentage score (0-100%) awarded on the last, comprehensive exam, which covers course material of the entire online course.

Methods of Computer-Mediated Communication: The communications options made available to the faculty and students by the software features of the online course management system (eCollege™ used by the university, which was the subject in the study.

Student Outcome: For purposes of this study, student outcome is defined as the level of participation of students in online courses and the final exam grades awarded to students upon completion of an online course.

Synchronous Communication: Online interactive or real-time communication between the student and instructor or among students.

Threaded Discussion: Asynchronous communication between students and instructor or among students, where participants post text-based messages.

Chapter 5.14
Exploring “Events” as an Information Systems Research Methodology

Anita Greenhill
The University of Manchester, UK

Gordon Fletcher
University of Salford, UK

ABSTRACT

In this article we build upon existing research and commentary from a variety of disciplinary sources, including information systems, organisational and management studies, and the social sciences that focus upon the meaning, significance and impact of “events” in the information technology, organisational and social context. Our aim is to define how the examination of the event is an appropriate, viable and useful information systems methodology. The line of argument we pursue is that by focusing on the “event” the researcher is able to more clearly observe and capture the complexity, multiplicity and mundaneity of everyday lived experience. An inherent danger of existing traditional “event” focused studies and “virtual” ethnographic approaches is the micromanagement of the research process. Using the notion of “event” has the potential to reduce methodological dilemmas such as this without effacing context (Peterson, 1998, p. 19). Similarly, in this article we address the overemphasis upon managerialist, structured and time-fixated praxis that is currently symptomatic of information systems research. All of these concerns are pivotal points of critique found within event-oriented literature regarding organisations (Gergen & Thatchenkery, 2004; Peterson, 1998).

INTRODUCTION

An examination of event-related theory within interpretative disciplines directs our focus toward the more specific realm of the “event scene.” The notion of the “event scene” originated in the action based (and antiacademy) imperatives of the
situationists and emerged in an academic sense as critical situational analysis. Event scenes are a focus for contemporary critical theory where they are utilised as a means of representing theoried inquiry in order to loosen the restrictions that historical and temporally bound analysis imposes upon most interpretative approaches. The use of event scenes as the framework for critiquing established conceptual assumptions is exemplified by their use in *CTheory*. In this journal’s version and articulation of the event-scene poetry, commentary, multivocal narrative and other techniques are legitimated as academic forms. These various forms of multidimensional and multivocal expression are drawn upon to enrich the understandings of the “event” to extricate its meaning and to provide a sense of the moment from which the point of analysis stems.

The objective of this paper is to advocate how information systems research can (or should) utilise an event scene oriented methodology. The paper is organised as follows: we begin by presenting the theoretical background and definitions of “event scenes” and the “event.” We do this as a means of illustrating how events capture multidimensional and multivocal forms of expression. The significance of this method is that it is a nonlineal and less time focused approach that has the potential to challenge the managerialist, structured and time-fixated praxis that is currently dominating information systems research and development. In the next section we illustrate why and how event oriented methods advocate including elements of illogical asemiosis of experience that eschews the application of management process and articulates arhythmic patterns of life, including political and cultural experience. We then argue there is a need to utilise consumption based approaches in information systems research away from traditional production-based systems understandings. Finally and most importantly, utilising an event-based focus in information systems can challenge existing constructs that perpetuate mainstream regimes of power by widening the boundary of what we understand as “the system.”

**WHAT ARE EVENTS AND EVENT SCENES?**

*The whole life of those societies in which modern conditions of production prevail presents itself as an immense accumulation of spectacles. All that once was directly lived has become mere representation.* (Debord, 1994, Thesis 1)

In this paper we present a sample of literature concerning event-oriented approaches, especially those inspired by the situationists, in order to consider the more specific representational issues found in the specific praxis of the “event scene.” We build upon Peterson’s (1998) literature review that offers a taxonomy of organisational events to develop a critical debate regarding the relationships of events to organisations. The event scene is the direct descendant to the situationism’s act of détournement, in which significant and insignificant elements of observations are isolated and inserted into new and unexpected contexts. Détournement is most readily explained with examples such as found art and the work of artists such as Tracey Emin that includes her Curriculum Vita (CV) presented as a framed piece and more recently an abusive text message sent to a fan. A majority of Emin’s work places the mundane in a formal environment in unexpected ways, forcing the viewer to (hopefully) reconsider their position and view the subject of the works in new ways. As a necessarily obtuse explanation of this tactic, Debord and Wolman (1956) describe détournement as being “less effective the more it approaches a rational reply” to the cultural situation it critiques. The situationist’s invocation for obscurity is a political resistance to the likelihood of mainstream recuperation — of being made ir-
Exploring “Events” as an Information Systems Research Methodology

relevant by becoming commodified. Event scenes are a mechanism utilised by contemporary critical theory in order to loosen the restrictions of historical and temporally bound analysis that are a consequence of most interpretive methods.

Our emphasis is primarily interpretative and contrasts with the growing use of Complex Event Processing (Mohamed, 2006; Niblett & Graham 2005). This theorisation of the event has developed from a computer science and processual perspective. While the founding rationale for this approach could arguably be seen as similar to our own, its implementation and general focus of attention differs significantly.

The general theoretical orientation of this work is drawn from the situationism of Debord’s (1994) Society of the Spectacle (Albright, 2003), de Certeau (1988), Lefebvre (1992) and a cautious reading of Baudrillard’s work (1988, 1998) regarding simulation and hyperreality. We acknowledge that this selection is a somewhat distorted representation of a situationist work. Debord and Baudrillard, for example, have been claimed as being at odds with one another intellectually at various points in their careers, as well as with the general development of situationism (Albright, 2003). Debord’s (1994) identification of the spectacle informs the meaning of the event scene used in this paper and assists in justifying our position that it is a legitimate approach to researching contemporary cultural and organisational phenomena. Baudrillard’s (1993) argument regarding the balance between the mundanity of everyday life and moments of tension also positions the observation of event-driven culture.

In contemporary culture, even for the Frankfurt School of Critical Studies two generations ago, “diversion, distraction, and amusement had become the norm.” (Hoover & Stokes, 1998). Attention to the minuscule of everyday life is both the norm of everyday life as well as being the representational tactic employed within the event scene (Peterson, 1998, p. 20).

The focal point of the event could be claimed as a complex potlatch; it is no coincidence that Potlatch was a key journal that inspired original situationist thought. The event is the mundane, the integrative blending of moments that constitute everyday life, the nonlinearity of experience, the illogic of expectations, the indeterminate acceleration and deceleration of personal temporality and the moments of the unexpected or unforeseen (Albright, 2003; Debord, 1994; Gergen & Thatchenkery, 2004; Peterson, 1998, p. 24; Plant, 1997, p. 236). The researcher is politically obliged within this framework to represent the event (any event) as an event scene — the excised moment of observation and experience captured and individually emphasised by them (Peterson, 1998, p. 20). Representing human experiences in the context of the “immediate” supports our claim that research methods that are less dependent upon historical and temporal references have the potential to reveal alternative and important understandings of information systems development and use.

EVENTS AND COLLECTIVE MEANING

Collectively considered together, “events” are the combination of situations and occurrences that have persistent significance to a social group as shared meaning-making and identity-making constructions (Urry, 2002). The examination and

In the face of the threats of a total weightlessness, an unbearable lightness of being, a universal promiscuity and a linearity of processes liable to plunge us into the void, the sudden whirlpools that we dub catastrophes are really the thing that saves us from catastrophe. Anomalies and aberrations of this kind recreate zones of gravity and density that counter dispersion. (Baudrillard, 1993, p. 69)
Exploring “Events” as an Information Systems Research Methodology

intermediate of events and their representation as event scenes is not a new research enterprise. Its foundations lie in classical historical analysis, including the documentation of significant moments of humanities’ progress through time (see Burke’s 1978 mainstream détournement). As Burke (1978) acknowledges through his own somewhat unconventional view of history, an event offers different meaning to different social groups that reflect divergent genealogies of events. Within information systems research, attention to structure and process produces a lack of sensitivity to the everyday and constant interplay of events. Plant (1997, p. 12) also provides an indefinite definitional basis for the spectacle when she identifies it as the “materialisation of ideology.” In the broadest sociological sense, contemporary events include reality television programmes, sports fixtures and annual festivals. All of these have been made the focus of theoried examination through a variety of methodological lenses. In an information systems context events exist in a variety of forms, including version change, system failure (in its many well-documented permutations), new personnel and new cohorts of “users” (such as the annual induction of higher education students to virtual learning environments). Events in this way are imprecisely situated within a historical, temporal, political and locational morass. The logic of information systems events is more clearly defined and understood by their shifting interrelationship to one another rather than their position on a Gantt chart, in a timeline, physical location or particular management regime.

It is those events that are shared and recalled (although not necessarily in any linear or logical fashion) as significant referents that engender cultural dynamism and contribute to the perpetuation of social structures. It has been argued by Urry (2002) that events such as wars, inventions, rituals, ceremonies, births and deaths are the core elements in the construction of shared meaning and are vital for the establishment of individual as well as social identity. The documentation of past events — or, alternatively, written histories — are significant cultural artefacts that retain collective consciousness in a tangible and objectified manner. Similarly, the documentation of future events in procedures and system designs embed historical, temporal, political and locational bias and assumptions that are effaced (or at least obscured) by the internal “logic” of documentation practice and the “structure” of a system’s design. These realisations are implied as central concerns for the situationists with their criticism of contemporary art and visual representation (Plant, 1997). Situationist thought, which by implication informs the event-driven perspective more generally, understands that the indirect experience of the event encapsulates a hidden but mediating representation that contributes to the obfuscation of the influences that the holders of “real” power have in contemporary society (Albright, 2003). The mechanism by which events or other units of enquiry are represented and labeled through a seemingly neutral “methodology” is consequently recognised as a powerful (and empowering) aspect of the research process. The embedded political relations found in research-based representations of events also contributes to a wider agenda that preserves the existing structures of mainstream power, whether this be political, gendered, ethnic or economic. De Certeau (1988, p. xvii) expresses this concern as the marginality of the majority:

Marginality is today no longer limited to minority, but is rather massive and pervasive: this cultural activity of the non-producers of culture, an activity that is unsigned, unreadable and unsymbolized, remains the only one possible for all those who nevertheless buy and pay for the showy products through which a productivist economy articulates itself.

The event that was the attack on New York in September 2001 and its later evolution into the media-driven event scenes of “9/11” a year later is one indication that the representation of
Exploring “Events” as an Information Systems Research Methodology

Events is a powerful political tool. The difficulty with representational strategies is that they can be used equally by situationists and critics of contemporary culture as well as by the holders of existing power (Albright, 2003; Plant, 1997). Realisation that the political motivations of situationism had itself been recuperated by the mainstream as “witty” ads and ironic play was a pivotal cause in the fracturing and dissolution of the movement. The Sex Pistols are one example of this tension. As their manager manipulated mainstream sensibilities to commercial success, the band’s own initial political and social commentary became increasingly questionable. We advocate, in the largely conservative environment of organisational studies, a critical re-examination of what methodology “does” but do not ad hoc reject all existing methodological paradigms (Gergen & Thatchenkery, 2004, p. 235). The issue being critiqued here is the current practice within information system’s research for continuous, but empty, justification and reiteration of “its” methods. Modernist desire for self-legitimation obscures recognition of the continuous sequence of interrelated events that is the information system in order to emphasise the research activity itself and to legitimate its utility (Gergen & Thatchenkery, 2004, p. 240). More significantly, debate concerning methodological appropriateness, if we apply the concerns of situationism, obscures examination of the real power holders who benefit from the events that are represented.

A VIEW OF THE EVERYDAY

To dérive was to notice the way in which certain areas, streets, or buildings resonate with states of mind, inclinations, and desires, and to seek out reasons for movement other than those for which an environment was designed. (Plant, 1997)

At first glance the effort to dérive (to become a derivite, to drift) appears to be the opposite political action called for by the desire to represent, recognise, and respond to the complacency of mundanity. However, the act of dérive is better viewed as the political and methodological act of looking beyond the veil of hegemonic expectations in order to see the actuality of use in places and with things. In an information systems context this could be (merely) seen as looking beyond the managerialist and structuralist views of a system (Peterson, 1998). The contemporary seminal example of the act of dérive are the unfocused, random and personal actions of the “Web surfer” (Andersen, 1998; Hartmann, 2003). Observation and participation within a system is contextual within a continuous sequence of interrelated events that captures what is actually done on a day to day basis rather than what is expected of individuals.

The role of the everyday within information systems research is, however, only marginally articulated or acknowledged in the majority of seminal information systems literature. Such a paucity of material is despite the significant impact that information systems themselves have upon daily life, both in a workplace context and increasingly in the domestic environment. However, the discussion of everyday life and its critical debates are well covered elsewhere (in other disciplinary contexts) by writers such as de Certeau (1988) and Lefebvre (1992). Both of these authors had also recognised association with situationist thought. De Certeau (1988, p. xviii) in discussing the personal interrelationships of everyday life claims that statistical investigation remains virtually ignorant of these trajectories, since it is satisfied with classifying, calculating and putting into tables the “lexical” units which compose them but to which they cannot be reduced, and with doing this in reference to its own categories and taxonomies.

These claims can be rightly construed as a critique of quantitative and positivist praxis.
The implication in a critique of this type is that these approaches produce their own internal logic that obscures external influences of power upon those people and “things” being tabulated, and who are ultimately affected. De Certeau (1988) argues for the significance of the interrelatedness of everyday life when he claims that “the analysis of the images broadcast by television and of the time spent watching television should be complemented by a study of what the cultural consumer ‘makes’ or ‘does’ during this time and with these images.” Basden (2005), in examining the works of Dooyeweerd, also claims that

though we cannot theorize scientifically about everyday life, we can understand it philosophically as an integration of the aspects of our experience. In the everyday, all aspects play their proper place to a greater or lesser extent. This is why, for example, it has an important social aspect and a religious (pistic) aspect, as well as a sensory aspect. But it also means that everyday living is not devoid of analytical activity (which is akin to theoretical thinking), though this takes the form of an analytical subject-object relationship rather than a theoretical Gegenstand-relation. This provides a useful foundation for analysing the richness of everyday life, everyday engaged attitudes and tacit knowledge.

The situationist view of the everyday is not, however, celebratory. Situationist association of everyday life with people’s oppression and disempowerment largely prohibits this perspective, at least directly. De Certeau’s general observations regarding everyday life also reflect this political hesitation.

As the indivisible stage of experience, everyday life is the venue for the construction and articulation of events (Peterson, 1998, p. 20). Understanding information systems in this context places them in the realm of everyday life, where they cease to exist in any systematic or singular sense. Information systems (whatever these may be) as an experience of everyday life become (perhaps merely) a surfeit of received information so that “today, the population is subjected to a continuous bombardment of damned stupidities that are not in the least dependent on the mass media” (Debord, 2002, p. 130). Debord (2002, p. 130) sustains this critique by claiming that “information theory straightaway ignores the chief power of language which lies on its poetic level; to compete and supercede.” Information does not have inherent power solely as a consequence of the scale of individual collections (with the Internet being the uber example) but in conjunction with the manner that information is read and reinterpreted; in short, how it is presented and represented.

THE MULTIPLICITY OF EXPERIENCE

The event-based approach is a rejection of the linearity of practice that is assumed within predominant “systems” based approaches. Methods that seek to understand “the system” commence with a series of assumptions that include the belief in an a priori presence of a system “merely” because the concentrated accumulation and representations of information is labeled as such.
Belief in the systemic nature of everyday life is an agreement of coherence and semiosis that is not borne out by experience. An event-oriented method, in contrast, integrates and recognises the illogical asemiosis of experience that eschews the application of management process and articulates arhythmic patterns of life (Peterson, 1998). For example, organisational studies of university management would attend to the application of documented policy, the process of committee based decision making and the general hierarchy of authority within a university. An event-oriented approach applied to the same environment may focus on the “hidden” flows of e-mails between colleagues, the use of “chair’s action” for decision making and the day to day solving of problems that contradict documented policy.

Critical and contemporary social studies no longer unwittingly accept the dominant historical accounts of events as the only “truth.” Nor do these perspectives blindly accept the unseen influences of power in the constitution of social phenomena or even identity. Writers such as Baudrillard (1998), Bergson (1910) and Game (1996) have all challenged the linearity of history and have gone so far as to argue for, at least in some cases, the death of history. In this rethinking of history the temporally cemented event — fixed, reified and glorified — is challenged. Raising doubt regarding the “certainty” of specifically identified events is particularly true in relation to the role that history plays as the central parameter of cultural understanding. The exploration of notions of time and space and their philosophical relationships is a central focus in contemporary studies of the social. Time, Bergson argues, has philosophically become spatialised (1910). That is, when time is spatialised it is understood as being able to be touched, seen as having discrete elements or presence and, most significantly, it can be presumed to be represented in this way (Game, 1996). As Game states “the common conception of time is that it is abstract, linear, and homogenous; homogenous empty time” (Game, 1996, p. 95). It is Bergson’s notion of multiplicity, however, that positions — even anchors — the construction of event scenes and the representation of everyday life experiences. The method of multiplicity, outlined by Bergson, employs dislocation by taking any object and disassociating its different moments, or its different ways of meaning (Game, 1996, p. 92). Bergson’s multiplicity mirrors the political strategy and research method of détournement (the disentanglement of cultural products to present new and oppositional meanings). Pulling apart the normality of everyday life plays with the meanings (and understandings) of the single instance and multiple events (their interlocking relationships) (Debord & Wolman, 1956). Bergson is critical of the approach and idea that “the present contains nothing more than the past, and what is found in the effect was already the cause” (Bergson, 1910, p. 15). The influences of historical materialism is readily identified within both qualitative ethnographic works and quantitative longitudinal studies (two methodological approaches that currently find favour within information systems research). Historical materialism presumes that what has preceded is the key relationship and source of understanding. However, for the contemporary moment it is an under theorised enquiry.

**CAPTURING THE EVENT WITHIN THE RESEARCH PROCESS**

The diverse approaches to systems thinking utilised in the field of information systems necessitates critical engagement with the question of the research position as the actual site of analysis and point of observation of the event. Systems theory methodologies require a boundary to be placed at the site of analysis (Heylighen, 1998). Encapsulating the subject under examination reduces the endless combinations and interactions of complex systems — the wide range of
Exploring “Events” as an Information Systems Research Methodology

events — that can be observed. It is at the nexus between advocacy for the need for boundaries and alternatively their permeability that debates regarding the meaning and purpose of information systems research exists. General systems theory applies boundaries in order to present a minimalised but holistic position of analysis (Heylighen, 1998). The bounded conceptualisation of a system exists within the broader continuum of system theory approaches and has been utilised across the information systems field, which ranges from the “hard system/cybernetic” approaches to “soft systems” (see Checkland & Howell, 1998 for a history of systems thinking in information systems development).

Systems operating within organisations are usually considered open in that there is recognition of the dynamic interaction of the system with the surrounding environment (Robbins & Barnwell, 1994). The system’s boundary serves to enclose internal operational elements from those external to the system and environmental conditions which may impact upon the system as a whole. However, the system’s boundary is permeable (Greenhill, 2002). Within the system, information is processually transformed from input to the output stage. Systems developers expect and plan for information taken into a system to be altered in predictable systematic ways up to the final point of output, of release from the system. Bundled with this initial assumption, the meanings and purposes associated with specific information are also fixed. Baskerville and Pries-Heje’s (1998) study on the management of knowledge capability and maturity in a small to medium size software development organisation is a pertinent example of the expectations of the systems developers in developing and maintaining fixed meanings within systems. The company Baskerville and Pries-Heje studied experienced difficulties developing organisational and Web-based systems. Assessment of the company’s situation was carried out primarily in managerial terms, rather than as a sequence of events, as the employees themselves claim.

I realise that all documents needed to support this, namely customer contract, project presentation, budget and requirements specification, were nowhere, and there were a thousand different meanings within Proventum about how they should look... Today we have as many different contracts as we have employees, because we don’t have a template to work from. (Jan in Baskerville and Pries-Heje, 1998, p. 183)

We need to be better at exploiting the knowledge from previous projects, much better, so we don’t make the same things again and again and again. (Henrik in Baskerville and Pries-Heje, 1998, p. 183)

Many systems designs rely on the fact that there can be no unsystematic, nonprocessual or unexpected alterations to the meaning of information (no unforeseen events). The experience of the system is not regarded as a varied combination of interconnected events but a continuous timeline of neatly packaged and logical actions. In the case of this often cited example, information was seen in management terms to remain static and continuously available to enable its exploitation in the future. However, from an event-based perspective the purpose and use of the information within a planned system does not necessitate or provide any singular or fixed accumulation of meanings. In place of uniform certainty is an array of interrelated meanings that the user may variously interpret from a system and its usage. It is only the genealogy and association crafted through organisational culture that produces mutually shared understanding of the system. As Wittgenstein observed, there can be no “private language.” Understanding and mutual comprehensibility is a joint-action (Gergen & Thatchenkery, 2004).

The interpretive position offered by Orlikowski & Gash (1994) and Feldman & March (1981) are
two examples of the influence of situationism on information systems research. The assertion that information always holds multiple meanings challenges any methodological assumption surrounding the construction and representation of meaning that presumes a linear monodimensional process (Baudrillard, 1994). An event-oriented perspective enables the identification of many taken-for-granted positions to be found in current methodological frameworks through the act of détournement, and reveals rather than obscures the political environment around which information is manipulated (Baskerville, Travis, & Truex, 1992). The foundational model for information systems was developed within a modernist context and utilises static and linear understanding to the meaning of information. It is only now that we, as participants within a postmodern cultural condition, are able to critique and question the appropriateness of static models of systems models usage.

Generally, analysis of systems operations and information systems utilise the modernist tradition, which emphasises and restricts understanding through the processes of production (see Alter, 1996; Hirscheim, Klein & Lyytinen, 1995). The information system is represented and understood as a Fordist information “factory.” For example, the system Baskerville and Pries-Heje (1998) explore is a knowledge management system that requires the system output – knowledge – to be managed. The input into the system is garnered from knowledge obtained from new and existing employees on the software programming team. More specifically, new knowledge, meanings and hence inputs that the individuals have in relation to database technology, Internet technology and Web technology are inserted into the system (Baskerville & Pries-Heje, 1998). The management of this knowledge within the monolithic managerial perspective requires information at the input stage to remain deterministically static in order to be both predictable and reusable. The application of this knowledge may be required at a later date (as systemic output). Therefore, it must be controlled in terms similar to those of a factory process. The system is valued purely in terms of this restrictive and narrow, but tangible, output. Processual approaches restrict the understanding of a system to an examination of data, its utilisation and manipulation. The goal or objective of the system is reached by asking whether the end product or output is effective and achieves the desired outcomes. Outcomes are generally assessed from a managerial perspective, simultaneously reducing the day-to-day user to a component of the system. Mechanistic positions reflect much of the contemporary information systems thinking (see Alter, 1996; Achterberg, van Es & Heng, 1991; Hirscheim, Klein & Lyytinen, 1995; Morgan & Smircich, 1980). Incorporating the event into systems analysis challenges production driven theorisations by shifting analysis away from production and the privilege of the manager. Consideration of events and the role of less privileged users reduces the dominance of hierarchical and managerial views of the system. Ultimately, the political aim of this perspective is to reposition those who contribute to events in the system as owners of that system.

**METHODS FOR REVEALING AND REPRESENTING THE EVENT**

Much of the methodological challenge that is taken up by event based analysis has been described ad hoc under the rubric – “the postmodern turn” (Brown, 1990, p. 196). Although this “turn” is yet to be fully articulated within the studies of organisations, commentaries such as that of Gergen and Thatchenkery (2004) do justify the turn for further critical work. Although postmodernism has been posited as antithetical to the modernist project, the relationship is not simply a structuralist dichotomy (Lagopoulos, 1993). Foster (1983, p. xi) suggests that the task of postmodernism is to extend the project of modernism by opening
Exploring “Events” as an Information Systems Research Methodology

“its closed systems to the heterogeneity of texts.” The politically confining aspects of modernist methodology are alternately accentuated, ignored or rejected in the various “postmodern” approaches (Huysen, 1992; Jencks, 1992.). Klotz (1988), in his attempt at reaching a definition of architectural postmodernism — itself a potentially modernist task — provides the basic framework for critical social research. He cites ten defining characteristics of the postmodern experience, ranging from geographic specificity and poetic cultural constructions to a need for relativism. Recognition of these cultural conditions irrevocably alters the justification of modernist methodology. The recognisable traits of postmodernism all emphasise the irrevocably altered nature of social relations in advanced mainstream capitalism, including what we claim is a surfeit of events. Among other qualities found in Klotz’s (1988) definition of postmodernism are the use of fiction in conjunction with function, the ironic “use” of history, the plurality of style, and a movement away from the perceived inevitability of technological progress. Capturing this complex social environment is more readily done through the gaze of the dérive — looking at the world of everyday life from the outside — and engaging in détournement — tearing down the supposed stability of systems.

The urban form, the visual, a celebration of the mundane, the embodiment of readable messages within material culture items and, obliquely, the increasing importance of entertainment in daily life are all elements of a critical methodology that attempts to understand contemporary social life. The movement away from modernist method and its quantifying concerns has been paralleled with an interest in the study of the popular and — to its extreme manifestation — kitsch (Jameson, 1983). What had been previously dismissed as not worthy of study or as being simply ugly have acquired undiscovered qualities, bringing them into the framework of theoried examination. The academic study of the products in the everyday life of mainstream capitalism, such as tourist’s souvenirs and the car, is compatible with the attitudes of the dérive.

Examining the mundane “things” of everyday life and their relationship to other “things” also emphasises a shift in focus from production-based analysis to more consumption-orientated approaches. This is a view which is confirmed by Shields (1992, p. 2), who believes that “in general, the modernist separation of economy and culture has left little room for serious engagement with consumption practices.” Consumption-based methods provide a degree of flexibility and encompass a significant part of an individual’s social life. Being “out” in the public sphere is to be consuming, not just foodstuffs and fuels, but more intangible items, including events and information. The practice of consuming in advanced capitalist social life has become synonymous with social participation (Derrida, 1978). People’s ability to remain social participants is determined by their consumption practices. In this sense consuming events and gazing upon objects are important aspects of everyday life, and by implication the research methodologies concerned with human experience. Consumption encompasses a significant proportion of social life when the supposedly “ordinary” can be viewed both as spectacle and as the parody of spectacle — the unspectacular event.

Within the context of everyday life the consumption of events is sublimated into the realm of the ordinary slipping from political consciousness to reinforce existing power structures. The situationists provide two methodological tactics that support their underlying theoretical and political standpoint: dérive and détournement. A third methodological tactic was identified by the situationists as the position and action that must be resisted, that of recuperation (being subsumed into the mainstream). Much of the obscurity, complexity and incoherence of the original situationist works was incorporated as a defense mechanism against this counter-tactic.
Criticism of later works with a situationist heritage could also be understood through the realisation of this tactic and proactive resistance. Writings such as Baudrillard’s (1998) “postmodernism” and Derrida’s (1978) “poststructuralism” are two immediately obvious examples. There is a tendency to isolate an individual tactic of situationism and celebrate its relevance. Of the three methods this response is most commonly found with the dérive and the Web (Andersen, 1998; Hartmann, 2003). However, this methodological isolation is a disservice to the original intent of the situationists. The *Internationale Situationiste* (1958) claim for détournement, as “the integration of past or present artistic production into a superior construction of a milieu,” continues to have relevance in relation to information systems as an inter-related combination of events that are not constrained by the arbitrary boundary of a documented system. Setting the information system free of unfounded delineation requires the act of the dérive. To drift and discern the location of these relevant but distanciated events requires the attitude that mirrors Baudelaire’s *flâneur* who is not constrained by the conventions that the recuperated information system seeks to sustain and perpetuate upon its hierarchically labeled and systemically controlled users.

Existing examples of event-oriented critiques of information systems are not readily found as printed documents. However, the increasing domestication of information technology acts as an enabling mechanism that brings event-based critique to the Web. These are generally visual and visualisation projects. The *Digital Landfill* project ([www.potatoland.org/landfill](http://www.potatoland.org/landfill)) takes images and texts from randomly selected Web sites and détournes them into a single image that seems to be “almost” meaningful. Similarly, the home page of [spaceless.com](http://spaceless.com) generates a random selection of images gathered from everyday life that appears to be a coherent collection. The semiotic obscurity of the resultant combination of images resists recuperation while standing as a critique of the vast asemiotic information system that is the Web. A more focused use of situationist tactics to develop a critique of an e-commerce system was the ongoing dispute between *etoys.com* and *etoys.com* (Stallabrass, 2003). *Etoy.com* became the vehicle from which “toy war” was launched. This “war” parodied the techniques of online business to détourn the meaning of e-commerce. The final result of this political engagement was the corporate failure and bankruptcy of *etoys.com*.

However, some Internet art projects reach beyond détournement and this is explained by Debord and Wolman’s definition of the tactic. “The distortions introduced in the détourned elements must be as simplified as possible, since the main impact of a détournement is directly related to the conscious or semiconscious recollection of the original contexts of the elements” (Debord & Wolman, 1956). This tactic is consequently not an anarchic free for all, but rather a considered and theoried technique that specifically endeavours to produce a political response to the observed world and status quo.

### THE DEATH OF HISTORY (OR THE DEATH OF CRITICAL INFORMATION SYSTEMS RESEARCH)

This article has presented the definitions of event scenes and the theoretical background regarding events. We provide an overview of existing research and commentary that focus upon the meaning, significance and impact of “events” in the information technology, organisational and social contexts. Peterson’s (1998) literature review provides initial guidance in revealing the possibilities for a taxonomy of organisational events. In this way, we have developed the foundations for a critical debate regarding the relationships of events to organisations.

The argument presented here has demonstrated how the “event” as a method can capture multi-
dimensional and multivocal forms of expression. We have shown how the examination of the event can form the basis for an appropriate, viable and useful information systems methodology. By focusing on the “event” the researcher or system designer can observe and capture the complexity, multiplicity and mundane of everyday lived experience. By utilising an “event” focus in IS research, we argue for the potential to reconstitute the mundane, the integrative blend of moments that constitute everyday life, the nonlinearity of experience, the illogic of expectations, the indeterminate acceleration and deceleration of personal temporality and the moments of the unexpected or unforeseen (Albright, 2003; Debord, 1994; Gergen & Thatchenkery, 2004; Peterson, 1998; Plant, 1997).

We have argued that utilising the event-oriented method, including elements of illogical asemiosis of experience, eschews the application of management process and articulates arhythmic patterns of life. The significance of the event-based approach to information systems development and research is that a nonlineal method challenges managerialist, structured and time-fixated praxis that currently dominate information systems research and development. The implication of our critique is that these existing popular approaches produce their own internal logic that obscures the influences of power on people and “things” located and ordered within the system. The influences of historical materialism can readily be identified in the preference for current research approaches. Historical materialism presumes that what has preceded is the key relationship and source of understanding — the event, in contrast, is currently an under theorised enquiry. Such theoretical foundation means that systems developers expect and plan for information taken into a system to be altered in predictable systematic ways up to the final point of output, where they are released from the control of the system. However, from an event-based perspective, the purpose and use of the information within a planned system does not necessitate or provide any singular or fixed accumulation of meanings. Consideration of events and the role of less privileged users reduces the dominance of hierarchical and managerial views of the system. Ultimately, the political aim of an event-oriented perspective is to make those in the system the owners of that system.

Finally, what is being presented in the event and the event-scene is not a metamethod for information systems research but an attempt to incorporate the complexities of everyday life and the subtleties of political meaning into the sterility of systemic systems thinking. Event-oriented perspectives in information systems offer the opportunity to engage in détournement for the purpose of both understanding existing environments and contributing to the development of future systems’ “architecture.” The active engagement in re-engineering echoes the situationist’s own town planning and architectural experiments (Sadler, 1998). Rebuilding the component parts found in the détournement produces new, unexpected and politically challenging approaches to the mundane of everyday life. Information systems research, in contrast, has been recuperated from its inception. Its methods, philosophy and advocacy continuously return to questions of business efficiency, process improvement and time management. As information systems become increasingly domesticated, the continuous and automatic reiteration of these perspectives without debate or critique will merely serve to perpetuate existing mainstream regimes of power.

REFERENCES


ABSTRACT

The horizontal and vertical dimensions of individualism and collectivism are an important characteristic of cultures. These dimensions have many implications for the ways in which individual learners use and respond to interactive technologies. This article reports on a study that investigated the impact of culture, specifically horizontal individualism (HI), vertical individualism (VI), horizontal collectivism (HC), and vertical collectivism (VC) on the effectiveness of technology mediated learning. Results indicate that the four dimensional patterns have differing effects on the use of TML communication capabilities, feelings of sense of community, satisfaction with the TML experience, perceived learning, and declarative knowledge acquisition.

INTRODUCTION

Culture is emerging as an important variable in the investigation of Technology Mediated Learning (TML) (Anakwe, Kessler, & Christensen, 1999; Collis, 1999; Gunawardena, Nolla, Wilson, Lopez-Islas, Ramirez-Angel, & Megchun-Alpizar, 2001; Salvatore, 2002). TML has been defined as “an environment in which the learner’s interactions with learning materials (e.g., readings, assignments, exercises), peers, and/or instructors are mediated through advanced information technologies” (Alavi & Leidner, 2001, p. 2). This
Culture's Impact on Technology Mediated Learning

This article extends the current examination of cultural influences on the effectiveness of TML by investigating the impact of one cultural dimension: individualism-collectivism at the individual level of analysis. This level of analysis is best conceptualized as “fluctuating pressures or tendencies” (Singelis, Triandis, Bhawuk, & Gelfand, 1995, p. 243), rather than as a distinguishing attribute (used in categorizing nations (Hofstede, 1980).

This research emphasizes the analysis of culture in a single learning setting, as opposed to cross-cultural studies that focus on the impact of culture across different settings (usually multiple geographic locations). One of the major differences in using a single learning setting and examining culture from an individual level of analysis is that the cultural dimensions being studied are not necessarily viewed as bipolar opposites (Triandis, 2004b). Rather, individuals can attribute characteristics of each dimension to themselves. That is, they can be bicultural, having both individualistic and collectivistic traits (Yamada & Singelis, 1995).

Calling for more holistic investigations of how TML can be used to improve the efficiency of delivery and effectiveness of learning outcomes, Alavi and Leidner (2001) stressed the importance of considering underlying psychological processes that are affected by TML. In their suggested framework, however, individual characteristics known to influence these learning processes were conspicuously absent. As Triandis (2003a) states, “cultural psychologists think of culture ‘in’ the person … there is no psychological process that is not shaped, to some extent, by culture” (p. 30). Therefore, from the perspective of individuals, learning in a TML environment should be influenced by their cultural inclination. Of the myriad cultural dimensions that have been identified (see, for example, Myers & Tan, 2002; Straub, Loch, Evaristo, Karahanna, & Srite, 2002), Individualism–collectivism has received the most attention (Triandis, 2004b), at both national (i.e., cross-cultural) and individual levels. We believe this dimension of culture is a critical factor in determining the perceived importance of various TML characteristics (asynchronous vs. synchronous communication tools), the perception of the TML environment itself (i.e., its social context), and as a predictor of learning outcomes. The following sections will expand upon the TML model developed by Alavi and Leidner (2001), and discuss the implications of the individualism/collectivism dimension in order to build support for this study’s hypotheses. This will be followed by a presentation of the study setting, methodology, results and discussion.

RESEARCH MODEL

Alavi and Leidner (2001) contend that to best examine the effectiveness of TML, the “mutual influence” of a constellation of variables need examination. They propose that instructional technique, coupled with the learning environment will impact underlying psychological learning processes of individual learners, which in turn will impact learning outcomes (satisfaction, perceived learning, and actual learning).

It is our belief that, beyond the characteristics described by Alavi and Leidner (2001), the TML ecosystem consists of idiosyncratic predilections of individuals participating in the learning context (e.g., individualistic and collectivistic traits), and use of the training technology (e.g., asynchronous and synchronous) communication capabilities. In addition, it is our contention that while TML may occur in isolation (e.g., computer-based training), learning outcomes improve from exposure to a social learning context in which learners interact with each other (Richardson & Swan, 2003), the instructor (Arbaugh, 2001; Picciano, 2002), and course content (Swan, Shea, Fredericksen, Pickett, Pelz, & Maher, 2000). Thus, our research model (see Figure 1) indicates that the individualistic-collectivistic dimension of culture affects the use of the Web-based communication technology in addition to perceptions of social presence (the
feeling of closeness with other learners), sense of community, and learning outcomes (learner satisfaction, perceived learning performance, and actual learning performance). The next section will provide a general description of online TML to establish the context in which culture is to be examined in this research.

**TML LEARNING CONTEXT**

Within a TML environment, interaction between learner, content, peers, and the instructor are vital to learning effectiveness (Moore, 2002). The mutual responses and social interaction enabled by TML have been suggested as the most natural way for people to learn (Hiltz, 1994). Recent research has found support for the critical role of each of these forms of interaction in creating an effective learning environment for both individuals and groups (Arbaugh, 2001; Gunawardena & Zittle, 1997; Hiltz, Zhang, & Turoff, forthcoming; Schmidt & Ford, 2003; Yi & Davis, 2003). However, the learning environment is not created by the technologies used, but instead within the environment enabled through the technology. That is, it is through ongoing interactions within a shared social context, that students will feel they are part of a learning community (Rovai, 2002b), rather than isolated individuals completing a course concurrently. This connectedness can only occur as a result of multiple interactions and active engagement within the learning environment (Gunawardena, 1995; Walther, 1995).

For TML, though, the key to these interactions is that learners choose to begin the process by actively seeking to interact through this medium. When focusing on how to design effective TML initiatives, it is critical that we better understand the impact that learners’ cultural beliefs have on the use of these communication technologies. Thus the remainder of this section will focus on the impact of individualism and collectivisms on the processes and outcomes of TML.

**ELEMENTS OF THE MODEL**

**Individualism-Collectivism**

Individualism implies that social behavior is established by personal goals and does not overlie the goals of the collective, while collectivism implies that the group is more important than the individual and the people in the group are inclined or ready to cooperate. It has been suggested that
a student’s “predisposition toward learning is systematically influenced by the individualistic or collectivistic values that dominate in a learner’s culture” (Chang & Lim, 2002 p. 88).

From the process perspective of TML, with its reliance on interaction, collaboration, and a social learning context for effective TML, individualistic tendencies that favor personal goals over the collective, self-sufficiency (Hofstede, 1980), and task orientation and focus (Trumbull, Rothstein-Fisch, & Greeneld, 2000) can create barriers to use of the communication capabilities of a TML environment and thus reduce students’ ability to take part in the creation of a social learning context. On the other hand, individualistic characteristics that rely more on words to interpret meaning (Hall, 1976), a greater competence for earning group membership (Triandis & Suh, 2002), and ability to establish non intimate, short-term relationships (Triandis & Suh, 2002), factor positively into forming the needed interactions in a TML. In the end, the individualistic tendency to have calculated interpersonal relationships, entering into a relationship if they have something to gain (Triandis, 2004a), may be the determining factor in using the capabilities of a TML to form a community of learners and receive the greatest benefit from the learning experience.

Likewise, collectivist characteristics can be seen to have both positive and negative influences on a learner’s approach to TML. The importance of group identity over individual identity (Chang & Lim, 2002), a focus on social obligations (Triandis, 1995), and acceptance of communal interpersonal relations (Triandis, 2004a), all provide support for the collaborative interactions necessary in a TML environment. However, for the collectivist, reliance on gestures, timing and facial expressions (Francesco & Gold, 1998), lack of skills for entering into new groups (Triandis & Suh, 2002), and a focus on more intimate and long-term relationships (Triandis & Suh, 2002) can also create barriers to the effective use of TML environments.

From the above summary, one could come to the conclusion that individualistic tendencies, due to emphasis on “task,” should ceteris paribus be positively associated with perceived and actual learning outcomes. However, the same tendencies that provide for the adequate attention and motivation necessary for positive learning outcomes may conflict with content delivery in an online TML environment. Therefore, it is hypothesized that these individualistic tendencies should in general be negatively associated with social presence, sense of community and satisfaction. A similar hypothesis is suggested with respect to the collectivistic findings. Certain collectivistic traits should increase the use of TML capabilities and the creation of a social learning context while others can hamper a person’s ultimate goal of effective learning. Thus, we propose to use the horizontal and vertical dimensions of individualism and collectivism (Triandis, 1995, p. 40) to further our analysis of this cultural dimension, with the expectation that the more specific analysis (or microanalysis) will help to refine our understanding of the impact of individualism-collectivism on TML.

Triandis (1995) argued that individualism and collectivism are multidimensional or polythetic constructs (Triandis & Gelfand, 1998), and recommended that the conceptualization of both individualism and collectivism be modified based on the additional attribute of the horizontal or vertical nature of a population. The horizontal aspect emphasizes the equality or similarity of people, whereas the vertical aspect accepts inequality among individuals and views differences among people positively. The horizontal-vertical distinction parallels Hofstede’s (1980) notion of power distance that is based on one’s level of acceptance of unequal power distribution; individuals aligned with the vertical dimension tend to be more accepting of this difference than those who are aligned more along the horizontal dimension. Support for the distinctiveness of these dimensions has been reported in the lit-
erature (Abraham, 1997; Chiou, 2001; Singelis et al., 1995; Soh & Leong, 2002; Thomas & Au, 2002; Triandis, 1995; Triandis & Gelfand, 1998). According to Triandis, “given the viability of the constructs, future researchers should focus on incorporating these dimensions into theory and research in social and organizational psychology” (Triandis & Gelfand, 1998, p. 125)

Research on differences among the four patterns of Horizontal Collectivist (HC), Horizontal Individualist (HI), Vertical Collectivist (VC), and Vertical Individualist (VI) has found:

Horizontal collectivists merge with in-groups (e.g., family, tribe, coworkers, nation); the well-being of their in-groups is important to them. However, they do not feel subordinate to their in-groups. In contrast, vertical collectivists submit to the norms of their in-groups and are even willing to sacrifice their personal identities for their in-groups. Horizontal individualists are characterized by seeking individuality rather than distinctiveness, that is, they tend to “do their own thing” and not to compare themselves with others. Lastly, vertical individualists are especially concerned with comparing themselves with others. They believe that competition is the law of nature, and they desire to win in all kinds of competitions. (Chiou, 2001, p. 668)

In addition to these generalized characteristics, research findings indicate that certain values are associated with the horizontal and vertical aspects of individualism and collectivism (Oishi, Schimmack, Diener, & Suh, 1998; Schwartz, 1994; Soh & Leong, 2002). These findings indicate that of Schwartz’s 10 values, power and achievement are associated with VI, self-direction with HI, benevolence with HC, and conformity and tradition with VC (Oishi et al., 1998).

Using this framework, we propose that the cultural-specific patterns of HI, VI, HC, and VC will affect TML in different ways (see Table 2). For each component of the research model, Table 2 indicates the effect of each pattern (HI, VI, HC, and VC) on that aspect of the model. For example,

<table>
<thead>
<tr>
<th>Findings</th>
<th>Use of Communication</th>
<th>SP</th>
<th>Sense of Community</th>
<th>Learning Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Satisfaction</td>
</tr>
<tr>
<td>Horizontal Individualism (Self-Direct</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Vertical Individualism (Power and Achi</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Horizontal Collectivism (Benevolence)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Vertical Collectivism (Conformity and T</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

Table 2. Proposed effects of HI, VI, HC, VC on research model
HI, VI, and VC are all expected to be negatively related to social presence, while HC is expected to be positively associated with social presence. Overall, the table indicates that while there are positive aspects of all four cultural patterns, the tendencies of HC individuals are best suited for the TML environment.

**HYPOTHESIS DEVELOPMENT**

**Use of Communication**

Because HI is associated with self-direction, we expect this to have a barrier-like effect on the use of TML communication. Self-directed learning would best be achieved in a CBT environment where interactions are between the learner and the content, with occasional interactions with an instructor. In the TML environment of our research setting, the learning process and outcomes are predicated on the interactions with content and instructors, but importantly, also among other learners. Thus, the need to achieve for VI will lead to the use of the communication capabilities when they form the foundation of the learning environment. With VC, we believe that their nature to conform will have them participating and interacting based on the structure of TML. Finally, we believe that HC benevolence and ability to merge with groups will allow them to quickly form group bonds, leading to early and repeated interactions using the only means available in TML, the communication tools. Therefore, we propose the following hypotheses:

H1: Cultural attributes of HI, VI, HC, and VC are associated with the use of TML communication capabilities.
H1a: HI will be negatively associated with use of communication.
H1b: VI, HC, and VC will be positively associated with use of communication

**Social Presence**

Social Presence (SP) is defined as a reflection of “the degree of salience of the other person in the interaction and the consequent salience of the interpersonal relationships” (Short, Williams & Christie, 1976, p. 65). Social presence is not a manifestation of static media characteristics; rather, social presence is a reflection of user perceptions that emerge over multiple interactions (Burke & Chidambaram, 1999; Walther, 1995). SP has been shown to be an important factor in the effectiveness of TML (Richardson & Swan, 2003; Shin, 2003; Youngjin & Alavi, 2001). Tu (2000), linking social presence theory with social learning theories, states that social presence “is required to enhance and foster online social interaction, which is the major vehicle of social learning” (p. 27). Past research has demonstrated the importance of cultural differences influencing perceptions of social presence (e.g., Gefen & Straub, 1997). There is additional evidence that social group membership influences the extent to which one perceives social presence (Bhappu, Griffith, & Northcraft, 1997). Recently, cultural differences (power distance and individualism-collectivism) were found to have a differential effect on the perceptions of website appropriateness, including that of social presence (Simon, 2001), and Gefen and Straub (2004) found social presence has an influence on trusting in e-commerce, especially with regards to feelings of benevolence. Taking direction from these studies, we argue that the cultural dimensions examined here will also have differential effects on perceptions of social presence.

**HV/INDIVIDUALISM — HV/COLLECTIVISM AND SOCIAL PRESENCE**

Social presence is in part contingent on the salience of others. Therefore, it is expected that an
individualistic perspective that values the self over others, coupled with a focus on tasks rather than people, will result in lower perceived social presence. In contrast, collectivist tendencies, which focus on people rather than tasks, would be expected to be associated with increases in perceived social presence. In the TML environment in which others are not seen but experienced via asynchronous text messages, it can be difficult to recognize the learners behind the messages while remaining focused on the task at hand. That is, individualism can lead to the view that asynchronous text messages read seriatim are akin to additional supportive material such as lecture notes, rather than as evidence of salient others. However, the predisposition toward group identity, focus on social obligations, and communal relationships of collectivism makes it more likely that collectivist learners will perceive the salience of the other behind the text message, leading to greater perceptions of social presence. Likewise, we expect that horizontal tendencies will lead to more positive perceptions of social presence. In contrast, those with vertical tendencies will describe fewer instances of social presence.

In the TML environment, feelings of equality and benevolence (HC) will allow for freer expression of ideas and thus foster more interactions among learners. As these interactions proliferate, opportunities to perceive the salience of the others in the conversation will also expand, thus increasing perceptions of social presence. On the other hand, emphasizing differences and self-direction (HI), power and achievement (VI), and conformity and tradition (VC) among learners can lead toward lurking or the withholding of viewpoints, with the effect of reducing interactions among learners and thus reduced perceptions of social presence. Taken together, this analysis suggests that HC should perceive higher levels of SP and the remaining groups will display lower levels of SP. Thus, the following hypotheses are stated:

H2: Cultural attributes of HI, VI, HC, and VC are associated with individuals’ perceptions of social presence.
H2a: HC will be positively associated with sense of social presence.
H2b: HI, VC, and VI will be negatively associated with sense of social presence.

**SENSE OF COMMUNITY**

McMillan and Chavis (1986, p. 9) define community as “a feeling that members have of belonging, a feeling that members matter to one another and to the group, and a shared faith that members’ needs will be met through their commitment to be together.” An important element in the formation of community within the context of distance education is interaction (Rheingold, 2000; Rovai, 2002b), and the inability to generate interactions among learners may create less effective online TML environments (Karuppam & Karuppum, 1999; Ueltschy, 2001). Rovai (2000) defined classroom community as consisting of four components: spirit, trust, interaction, and learning. Of particular importance with regard to the cultural dimensions being studied is spirit which denotes “recognition of membership in a community and the feelings of friendship and cohesion” (p. 287), characteristics associated with collectivism. We expect, then, that similar to perceptions of social presence, sense of community will also be impacted by the horizontal and vertical aspects of individualism-collectivism.

Belenky, Clinchy, Goldberger, and Tarule (1986) theorized that patterns in textual communication can enhance or inhibit the formation of classroom community. They conjectured that text which is viewed as separate, autonomous, or independent will inhibit community formation whereas text viewed as being relational and connected enables community formation. In their
theory they classify these distinctive voices as being respectively, male and female dominated. Research by Herring (1996) and Blum (1999) supports this claim. Similarly, we propose that the distinction between individualism and collectivism can also be expected to influence text voice, and thus impact the sense of community.

**HV/INDIVIDUALISM — HV/COLLECTIVISM AND SENSE OF COMMUNITY**

Similar to social presence, a key factor in building a sense of community among learners is interaction. It seems apparent, based on the definition of sense of community given by McMillan and Chavis (1986), that the attributes associated with collectivism — importance of the group and relationships within the group — are more conducive to contributing to a sense of community within the learning context. Indeed, if the individualistic attribute of autonomy becomes dominant, a sense of community may not develop leading to a less effective online TML environment. Likewise, we expect that those persons with horizontal tendencies, emphasizing equality of people, will have more positive perceptions of sense of community than would be associated with vertical tendencies, with its emphasis on differences among people. We contend that the properties of benevolence (HC) and conformity (VC) will lead to a positive sense of community within the TML environment. Alternatively, emphasis on power and achievement (VI) and self-direction (HI) will lead to a negative sense of community. Thus, the following hypotheses are stated:

**H3:** Cultural attributes of HI, VI, HC, and VC are associated with individuals’ perceptions of sense of community.

**H3a:** HI and VI will be positively associated with sense of community.

**H3b:** HI and VI will be negatively associated with sense of community.

**HV/INDIVIDUALISM — HV/COLLECTIVISM AND LEARNING OUTCOMES**

Chang and Lim (2002), using social interdependence theory (Johnson & Johnson, 1989), found that an individualist’s focus on task helps to promote the reasoning process, whereas the collectivist’s focus on relationships facilitates the social and response process of learning. Since learning in a TML environment necessitates group participation and communication, it is expected that collectivism will be positively associated with satisfaction. However, since individualists are less concerned about group harmony and success, they will be less satisfied by the need to interact with the group in the social and response process of learning. With its focus on task and achievement, it is also expected that individualism will be positively associated with the perceived and actual learning outcomes. With its focus on the group, and the importance of relationships, collectivism should be positively associated with perceived and actual learning, but only to the extent that the social and response process improve an individual’s ability to learn the material. Further, we expect that horizontal tendencies will be associated with greater satisfaction with the process of learning in a TML environment, compared to vertical tendencies. Likewise, the vertical tendency to view people as different may create disadvantages in a TML environment by creating instances where individual knowledge becomes subservient to the group; in this case, circumstances of perceived learning and actual learning might be expected to be lower than in the horizontal dimension. Thus, we believe that aspects of HI, VI, and HC can be expected to have a positive effect on learning outcomes (both perceived and actual). However,
we believe VC’s tendency to submit to the norms of their in-groups coupled with a willingness to sacrifice their personal identities will negatively influence learning outcomes. Thus, the following hypotheses are stated:

H4: Cultural attributes of HI, VI, HC, and VC are associated with individuals’ learning outcomes (satisfaction, perceived learning, and actual learning).
H4a: HI and HC will be positively associated with all learning outcomes.
H4b: VI will be negatively associated with satisfaction but positively associated with perceived and actual learning.
H4c: VC will be positively associated with satisfaction but negatively associated with perceived and actual learning.

We believe that as learners’ awareness of social presence and sense of community increase, the learning environment becomes akin to a community of practice. In such a community of practice, robust knowledge and understandings are socially constructed through talk, activity, and interaction around meaningful problems and tools (Vygotsky, 1978). In other words, a culture of learning develops. In this community of practice, within the context of an online TML environment, students distribute their intellectual activity so that the burden of managing the process does not fall to any one individual and learning effectiveness is enhanced (Bransford, Brown, & Cocking, 1999). A review of the literature on social interaction in a computer supported environment (Kreijns, Kirschner, & Jochems, 2003) indicates a strong relationship between the social context of the TML environment and learning effectiveness (Hiltz, 1994). Thus we hypothesize that:

H5: Greater perceptions of social presence will be associated with greater learning effectiveness outcomes (learner satisfaction, perceived learning performance, and actual learning performance).
H6: Greater perceptions of sense of community will be associated with greater learning effectiveness outcomes (learner satisfaction, perceived learning performance, and actual learning performance).

Research Setting and Procedure

Participants in this study are from an introductory IS course taught at a university in the southeastern United States. The course is required of all business majors and is taught entirely online, with no face-to-face sections taught during the semester in which the data was collected (Spring 2004). The course was taught using the WebCT platform. Approximately 1100 students enroll in this course each semester. To effectively manage such a large number of students, groups of approximately 35 are created and the students in these groups interact only within the group. The course is taught by one instructor with multiple teaching assistants. The course is set up as six independent modules in which the students have two weeks to complete the material. Each module focuses on different topical areas such as the strategic use of information and technology, e-commerce, and decision support systems. Within each module students are required to make a minimum number of postings utilizing the asynchronous capabilities of WebCT. At the end of each module an examination is given and students are assessed based on the total of the five highest module grades. Students are not allowed to proceed to the next module until the established opening date and time for the module. In other words student have relatively little control over some aspects of the learning process.

The course was designed around the communication tools available in WebCT, which included threaded discussions, course e-mail, and online chat. The discussion capabilities of WebCT were
Culture’s Impact on Technology Mediated Learning

the backbone for the course. For each of the six course modules, learners were required to post their answers to assigned questions from an end of chapter case and to respond to at least one other student’s answers. Although chat and e-mail capabilities were available, their use was not required. Chat capabilities were rarely used.

A survey to capture the elements of the model was delivered via WebCT as part of the first and fourth Module. Demographic data and cultural items related to HI, VI, HC, and VC were collected during the first module of the course. During the fourth module, questions regarding perceptions of social presence, sense of community, satisfaction and perceived learning were gathered. Actual learning was assessed from the grade received as a result of the end of module examination.

The study variables were measured as follows: Perceived Satisfaction was measured with a 7-item instrument developed by Biner (1993). Perceived performance was measured using the 12-item instrument utilized by Alavi (1994), and actual performance was measured using the results from the first three examinations. Social presence is measured with a 5-item instrument developed by Short, Williams and Christie, (1976). HI, VI, HC, and VC was measured with a 32-item instrument developed by Singelis et al. (1995). Sense of community was measured using a 10-item instrument developed by Rovai (2002a). To measure the use of the Web-based technology, we focused on asynchronous discussions as this was the primary means of communication within the course. Logs were provided in which we were able to tally the number of posts read, the number of posts made and the number of follow-up posts (a response to other students’ posts). The actual numbers of posts were used in our analysis. We combined these three values to form an overall measure of communication use.

Students were given extra-credit towards their final course grade for participating in this research. 524 usable results were obtained. Respondents included 229 men and 295 women.

Analysis: Measurement Model

Analysis was conducted using Partial Least Squares (PLSGraph 3.0) with scales measured as reflective, with the exception of communication use and actual performance. Several tests were performed on the measurement model to test for validity and reliability. Table 3 includes loadings of the reflective measures on their respective constructs along with composite reliability scores, standard errors, and t-statistics. It also includes the weights of the formative items. All reflective items are significant at the .99 level and all have high loadings. Most are above the 0.70 threshold, though the few loadings below this threshold are above the 0.50 suggested by Fornell and Larcker (1987), demonstrating convergent validity. The composite reliability scores of the latent constructs are all above the recommended 0.80 (Nunnally, 1978), demonstrating internal consistency.

Table 4 presents the correlations among the constructs and provides discriminant validity statistics. The square roots of the AVE scores (diagonal elements of Table 4) are greater than the correlations among the constructs, demonstrating discriminant validity. Furthermore, all items loaded higher on their respective constructs than on others, providing additional support for discriminant validity.

Structural Model

The results of the tests of the structural model and associated path coefficients are found in Table 5. Consistent with previous research and recommendations (Chin, 1998), bootstrapping (100 subsamples) was performed to determine the statistical significance of each path coefficient using t-tests. Results indicated that social presence, sense of community, communication utilization, and the cultural dimensions of HI, VI, HC, and VC were related to the TML outcomes. Each of the individualism-collectivism dimensions were related to at least one of the TML process constructs of
**Table 3. Measurement model**

<table>
<thead>
<tr>
<th>Construct</th>
<th>Items</th>
<th>Loadings</th>
<th>Weights</th>
<th>Standard Error</th>
<th>t-statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perceived Learning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = .936$</td>
<td>PL4 &amp; PL5 &amp; PL6 &amp; PL7 &amp; PL8 &amp; PL9 &amp; PL10 &amp; PL11</td>
<td>0.7565 &amp; 0.8128 &amp; 0.8628 &amp; 0.8535 &amp; 0.8153 &amp; 0.7873 &amp; 0.7697 &amp; 0.7642</td>
<td>0.00294 &amp; 0.00264 &amp; 0.0184 &amp; 0.0157 &amp; 0.0188 &amp; 0.0206 &amp; 0.0257 &amp; 0.0276</td>
<td>25.7691</td>
<td>30.7569</td>
</tr>
<tr>
<td><strong>Social Presence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.855$</td>
<td>SP1 &amp; SP2 &amp; SP3 &amp; SP4 &amp; SP5</td>
<td>0.7067 &amp; 0.8028 &amp; 0.7001 &amp; 0.8043 &amp; 0.6578</td>
<td>0.00318 &amp; 0.00197 &amp; 0.00331 &amp; 0.00190 &amp; 0.00336</td>
<td>22.2494</td>
<td>40.6643</td>
</tr>
<tr>
<td><strong>Sense of Community</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.888$</td>
<td>SC3 &amp; SC5 &amp; SC7 &amp; SC11 &amp; SC13 &amp; SC19</td>
<td>0.7733 &amp; 0.7537 &amp; 0.7324 &amp; 0.7214 &amp; 0.7746 &amp; 0.7700</td>
<td>0.00204 &amp; 0.00281 &amp; 0.00274 &amp; 0.00296 &amp; 0.00235 &amp; 0.00225</td>
<td>37.9735</td>
<td>26.8664</td>
</tr>
<tr>
<td><strong>Satisfaction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.878$</td>
<td>SAT3 &amp; SAT4 &amp; SAT5 &amp; SAT6</td>
<td>0.8171 &amp; 0.8273 &amp; 0.7591 &amp; 0.8014</td>
<td>0.00183 &amp; 0.00183 &amp; 0.00291 &amp; 0.00231</td>
<td>44.7252</td>
<td>45.1325</td>
</tr>
<tr>
<td><strong>HI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.820$</td>
<td>HI4 &amp; HI5 &amp; HI8</td>
<td>0.6517 &amp; 0.8443 &amp; 0.8244</td>
<td>0.00798 &amp; 0.00446 &amp; 0.00404</td>
<td>8.1635</td>
<td>18.9351</td>
</tr>
<tr>
<td><strong>VI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.768$</td>
<td>VI2 &amp; VI5 &amp; VI7</td>
<td>0.7020 &amp; 0.7859 &amp; 0.6836</td>
<td>0.2079 &amp; 0.1962 &amp; 0.1722</td>
<td>3.3772</td>
<td>4.0060</td>
</tr>
<tr>
<td><strong>HC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.837$</td>
<td>HC1 &amp; HC2 &amp; HC4 &amp; HC6</td>
<td>0.7687 &amp; 0.7690 &amp; 0.7216 &amp; 0.7377</td>
<td>0.0407 &amp; 0.0379 &amp; 0.0410 &amp; 0.0426</td>
<td>18.8879</td>
<td>20.2974</td>
</tr>
<tr>
<td><strong>VC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.808$</td>
<td>VC1 &amp; VC2 &amp; VC3</td>
<td>0.8285 &amp; 0.7844 &amp; 0.6734</td>
<td>0.0659 &amp; 0.0692 &amp; 0.1009</td>
<td>12.5697</td>
<td>11.3429</td>
</tr>
<tr>
<td><strong>Formative Measures:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Declarative Knowledge</strong></td>
<td>Quiz1 &amp; Quiz2 &amp; Quiz3</td>
<td>0.7760 &amp; 0.6513 &amp; 0.5250</td>
<td>0.1051 &amp; 0.1336 &amp; 0.0999</td>
<td>7.3801</td>
<td>4.8741</td>
</tr>
<tr>
<td><strong>Use</strong></td>
<td>Articles Read &amp; Articles Posted &amp; Follow ups</td>
<td>0.7847 &amp; 0.6563 &amp; 0.8234</td>
<td>0.0992 &amp; 0.1567 &amp; 0.0832</td>
<td>7.9071</td>
<td>4.1889</td>
</tr>
</tbody>
</table>
Table 4. Correlations between latent constructs

<table>
<thead>
<tr>
<th></th>
<th>PL</th>
<th>SP</th>
<th>DK</th>
<th>SC</th>
<th>Use</th>
<th>Sat</th>
<th>HI</th>
<th>VI</th>
<th>HC</th>
<th>VC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Learning</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Presence</td>
<td>0.38</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Declarative Knowledge</td>
<td>0.08</td>
<td>0.12</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sense of Community</td>
<td>0.31</td>
<td>0.52</td>
<td>0.02</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use</td>
<td>0.17</td>
<td>0.10</td>
<td>0.26</td>
<td>0.26</td>
<td>0.06</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfaction</td>
<td>0.43</td>
<td>0.47</td>
<td>0.22</td>
<td>0.31</td>
<td>0.26</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HI</td>
<td>0.17</td>
<td>0.04</td>
<td>-0.10</td>
<td>0.03</td>
<td>-0.07</td>
<td>-0.01</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>0.03</td>
<td>0.00</td>
<td>-0.07</td>
<td>0.07</td>
<td>-0.09</td>
<td>-0.11</td>
<td>0.16</td>
<td>0.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>0.30</td>
<td>0.14</td>
<td>0.03</td>
<td>0.23</td>
<td>0.10</td>
<td>0.21</td>
<td>0.21</td>
<td>-0.05</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>VC</td>
<td>0.10</td>
<td>0.06</td>
<td>0.07</td>
<td>0.15</td>
<td>0.04</td>
<td>0.08</td>
<td>0.02</td>
<td>0.06</td>
<td>0.19</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Diagonal elements are square roots of the AVE

Table 5. Structural model path coefficients, t-statistics and $R^2$

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>$\beta$</th>
<th>T</th>
<th>$R^2$</th>
<th>Dependent Variable</th>
<th>$\beta$</th>
<th>T</th>
<th>$R^2$</th>
<th>Dependent Variable</th>
<th>$\beta$</th>
<th>t</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use</td>
<td>-0.081</td>
<td>-1.40</td>
<td>.024</td>
<td>Social Presence</td>
<td>0.12</td>
<td>2.5</td>
<td>.021</td>
<td>Community</td>
<td>-0.029</td>
<td>5.5</td>
<td>.069</td>
</tr>
<tr>
<td>HI</td>
<td>-0.711</td>
<td>-1.37</td>
<td></td>
<td>VI</td>
<td>0.01</td>
<td>.02</td>
<td></td>
<td>HI</td>
<td>0.073</td>
<td>1.19</td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>2.12</td>
<td>2.12</td>
<td>*</td>
<td>HC</td>
<td>0.132</td>
<td>2.49****</td>
<td>Community</td>
<td>4.43****</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>0.032</td>
<td>0.46</td>
<td></td>
<td>VC</td>
<td>0.035</td>
<td>.76</td>
<td></td>
<td>VC</td>
<td>1.07</td>
<td>2.53****</td>
<td></td>
</tr>
<tr>
<td>VC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pcvd. Performance</td>
<td>0.239</td>
<td></td>
<td>.292</td>
<td>Decl. Knowledge</td>
<td>0.161</td>
<td>3.17****</td>
<td>Satisfaction</td>
<td>9.24****</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Presence</td>
<td>0.282</td>
<td>4.56****</td>
<td></td>
<td>Community</td>
<td>-0.124</td>
<td>1.84**</td>
<td>Community</td>
<td>1.29*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use</td>
<td>0.998</td>
<td>3.23****</td>
<td></td>
<td>Use</td>
<td>0.234</td>
<td>5.29****</td>
<td>Use</td>
<td>5.42****</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HI</td>
<td>0.125</td>
<td>2.79****</td>
<td></td>
<td>HI</td>
<td>-0.09</td>
<td>1.725**</td>
<td>HI</td>
<td></td>
<td>-0.021</td>
<td>.52</td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>0.116</td>
<td>2.79****</td>
<td></td>
<td>VI</td>
<td>-0.26</td>
<td>.51</td>
<td></td>
<td>VI</td>
<td>-0.085</td>
<td>1.91**</td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>0.023</td>
<td>0.52</td>
<td></td>
<td>HC</td>
<td>0.019</td>
<td>.39</td>
<td></td>
<td>HC</td>
<td>1.113</td>
<td>2.73****</td>
<td></td>
</tr>
<tr>
<td>VC</td>
<td>0.194</td>
<td>4.15****</td>
<td></td>
<td>VC</td>
<td>0.072</td>
<td>1.27</td>
<td></td>
<td>VC</td>
<td>0.023</td>
<td>.57</td>
<td></td>
</tr>
</tbody>
</table>

* $p \leq .10$ ** $p \leq .05$ *** $p \leq .025$ **** $p \leq .01$ ***** $p \leq .005$ (one tailed)

SP, sense of community or communication use. HC (.132, $p \leq .01$) was related to perceived social presence and explained 2% of its variance. HI (.217, $p \leq .005$) and VC (.107, $p \leq .01$) were related to sense of community and explained 7% of its variance. HI (-.081, $p \leq .10$) and VI (-.071, $p \leq .10$) were negatively related to communication use, whereas HC (.109, $p \leq .025$) was positively
related to use of communication tools; together they explained 2.4% of its variance. Social presence (.282, p ≤ .005), sense of community (.098, p ≤ .025), communication use (.125, p ≤ .005), HI (.116, p ≤ .005), and HC (.194, p ≤ .005) were positively related to perceived learning, explaining 24% of its variance. Social presence (.161, p ≤ .005), and use of communication tools (.234, p ≤ .005) were positively related to declarative knowledge, whereas sense of community (-.124, p ≤ .05) and HI (-.09, p ≤ .05) were negatively related to declarative knowledge; together they explained 10% of its variance. Social presence (.399, p ≤ .005), sense of community (.066, p ≤ .10), communication use (.193, p ≤ .005), and HC (.113, p ≤ .005) were positively related to satisfaction, whereas VI (-.085, p ≤ .05) was negatively related to satisfaction; together they explained 29% of its variance.

These results provide support for the research hypothesis that the cultural patterns of HI, VI, HC, and VC are related to various aspects of TML. In particular these cultural patterns have a large impact on TML outcomes (satisfaction, perceived learning, and declarative knowledge) explaining upwards of 30% of the variance compared with the emergent processes of TML (perceived SP, sense of community, and communication use) where upwards of 7% of the variance is explained. The next section will provide some insights and implications regarding the unbalanced impact that HI, VI, HC, and VC have on TML processes and outcomes.

**Discussion**

The results of this study provide overall support for the proposed research model. As predicted, the cultural patterns (HI, VI, HC, or VC) affected the learning outcomes in a TML environment, although not uniformly. Interestingly the attribute HI was the only cultural value to be associated with actual learning (although it appears to have had a negative impact). This suggests that HI individuals’ strong tendency to favor self-directed learning was not suitable for a TML environment in which collaboration plays a significant part of the learning process. With respect to perceived learning, the horizontal aspect turns out to be an important factor. Consequently, we can conclude that values associated with the horizontal aspect impact how persons perceive their performance in an online TML environment. Regarding satisfaction with the learning environment, our findings show that the attributes associated with HC, well being of their in-groups (presumably students from the class became their in-group), cooperation, and concordance with group members, contributed to a positive outlook of TML; whereas VI attributes, such as power and achievement, led to a negative view of TML, as the collaborative nature of the environment was not structured to support an individual but rather the interactions among individuals.

The two social context variables, SP and sense of community, were influenced by the collectivist dimension, with HC impacting perceived social presence while HC and VC both had an impact on sense of community. The importance of this finding underscores the effect that SP and sense of community has on TML outcomes. In TML environments in which interaction and collaboration are integral to learning outcomes, persons’ collectivist tendencies need to be nurtured for the positive effects they have on the processes associated with better TML outcomes. In contrast, the finding that the individualist dimension had no effect on either of these social context variables is an indication that an important structure in the learning process (i.e. the social context) may be inhibited as a result of a focus on individuality and self-direction, creating a situation where the connections necessary for a social learning environment are absent.

The attributes associated with HC, harmony of the group and cooperation, led to a strong association with communication use. Interestingly, both HI and VI with their tendency to value the
self over the collective led to lower levels of communication use. Pedagogically, by encouraging the characteristics associated with HC, we expect to indirectly increase the effectiveness of TML by increasing use of TML communication capabilities and perceptions of social presence and sense of community. Similarly, we would expect to directly increase the effectiveness of TML by simultaneously encouraging the characteristics associated with HC and HI and discouraging those associated with VI. Priming procedures have been effectively used to induce individualistic or collectivistic responses in previous research (Mandel, 2003; Trafimow, Triandis & Goto, 1991) and thus might prove an effective method for stimulating the positive attributes of HC within a TML environment.

Our results indicate that if HC characteristics can be positively encouraged, we can increase the intermediate TML processes of communication use (i.e., interaction) and the formation of a social context for learning in a TML environment to take place (SP and sense of community). We would also expect that in TML environments where interaction and collaboration are central to learning effectiveness that the ability to enhance HC characteristics will also lead to increases in satisfaction with the learning environment, perceived learning, and the extent to which declarative content is learned.

While the four patterns related to individualism and collectivism have been found to be associated to varying degrees with the effectiveness of TML, more research is needed. Other aspects of the TML environment might be affected by these culture patterns. Other dimensions of culture may also play a role. For example, Hall’s (1976) High and Low Context dimension might be very useful in understanding how individuals utilize TML communication capabilities, which in turn should both directly and indirectly impact learning outcomes.

Limitations

While the results of this study provide insights into designing effective TML initiatives, a number of limitations must be considered when interpreting the results. First, this study should be subjected to further testing with different participants, contexts, and technological architectures. Second, the research participants were undergraduate students who were completing the course as part of a degree requirement, so the results may not generalize to other settings and contexts. The cultural dimension of individualism-collectivism is but one individual cultural aspect that can affect TML process and outcomes, thus the conclusions drawn from this study should be done so cautiously as other influences may be as or more important to TML effectiveness.

Conclusion

This study has found support for the dimensions of horizontal and vertical individualism-collectivism as proposed by Triandis (1995). It extends the work of cross-cultural TML studies to investigate the impact that individual cultural beliefs have on the processes and outcomes of TML. We found that HC is positively associated with the process of social context formation — perceptions of social presence and sense of community, use of the communication tools, and with the outcomes of TML. Strengthening various characteristics of a person’s personality can lead to more effective TML initiatives.

ACKNOWLEDGMENT

Dr. Hornik would like to acknowledge the financial support provided by the Academy for Teaching, Learning and Leadership at the University of Central Florida.
REFERENCES


**ENDNOTE**

1 The data for this study was gathered as part of the second author’s undergraduate honors thesis which dictated when the final data collection could take place. As a result, we were unable to gather learning outcomes at the conclusion of the course.
APPENDIX A. RESEARCH INSTRUMENT ITEMS

Horizontal Individualism
1. I often do “my own thing”
2. One should live one’s life independently of others
3. I like my privacy
4. I prefer to be direct and forthright when discussing with people
5. I am a unique individual
6. What happens to me is my own doing
7. When I succeed, it is usually because of my abilities
8. I enjoy being unique and different from others in many ways

Vertical Individualism
1. It annoys me when other people perform better than I do
2. Competition is the law of nature
3. When another person does better than I do, I get tense and aroused
4. Without competition, it is not possible to have a good society
5. Winning is everything
6. It is important that I do my job better than others
7. I enjoy working in situations involving competition with others
8. Some people emphasize winning; I’m not one of them

Horizontal Collectivism
1. The well-being of my co-workers is important to me
2. If a co-worker gets a prize, I would feel proud
3. If a relative were in financial difficulty, I would help within my means
4. It is important to maintain harmony within my group
5. I like sharing little things with my neighbors

Vertical Collectivism
1. I would sacrifice an activity that I enjoy very much if my family did not approve of it
2. I would do what would please my family, even if I detested that activity
3. Before taking a major trip, I consult with most members of my family and many friends
4. I usually sacrifice my self-interest for the benefit of group
5. Children should be taught to place duty before pleasure
6. I hate to disagree with others in my group
7. We should keep our aging parents with us at home
8. Children should feel honored if their parents receive a distinguished award

Sense of Community
1. I feel that students in this course care about each other
2. I feel connected to others in this course
3. I do not feel a spirit of community
4. I feel that this course is like a family
5. I feel isolated in this course
6. I trust others in this course
7. I feel that I can rely on others in this course
8. I feel that members of this course depend on me
9. I feel uncertain about others in this course
10. I feel confident that others will support me

Social Presence
Higher numbers represent more presence and lower numbers represent less presence.
1. Impersonal...Personal
Culture’s Impact on Technology Mediated Learning

2. Unsociable...Sociable
3. Insensitive...Sensitive
4. Cold...Warm
5. Passive...Active

Satisfaction
1. I am satisfied with the clarity with which the class assignments were communicated
2. I am satisfied with the timeliness with which articles, tests, and written assignments were graded and returned
3. I am satisfied with the extent to which the instructor made the students feel that they were part of the class and “belonged”
4. I am satisfied with the instructor’s communication skills
5. I am satisfied with the in-person/e-mail/telephone accessibility of the instructor outside of class
6. I am satisfied with the present means of material exchange between myself and the course instructor
7. I am satisfied with the accessibility of the lab instructor

Communication
1. To what extent did you utilize chat for student-to-student communication?
2. To what extent did you utilize chat for student-to-instructor communication?
3. To what extent did you use chat for general course questions?
4. To what extent did you utilize chat for help on assignments?
5. To what extent did you use chat for assistance with technical issues (e.g., server down, WebCT problems)?
6. These next set of questions focus on your use of WebCT discussions.
7. To what extent did you utilize discussions for student-to-student communication?
8. To what extent did you utilize discussions for student-to-instructor communication?
9. To what extent did you use discussions for general course questions?
10. To what extent did you utilize discussions for help on assignments?
11. To what extent did you use discussions for assistance with technical issues (e.g., server down, WebCT problems)?

Perceived Learning
1. I did additional reading beyond the required assignments.
2. I did some thinking for myself.
3. I discussed topics outside the class.
4. I feel more confident in expressing ideas related to Information Technology.
5. I gained a good understanding of the basic concepts of the material.
6. I improved my ability to critically think about Information Technology.
7. I improved my ability to integrate facts and develop generalizations from the course material.
8. I increased my ability to critically analyze issues.
9. I learned a great deal of factual material in the course.
10. I learned to identify the central issues of the course.
11. I learned to interrelate the important issues in the course material.
12. I learned to value other points of view.

Chapter 5.16
Using Social Networking Analysis to Facilitate Knowledge Sharing in the British Council

Bonnie Wai-yi Cheuk
Improvement Service for the Scottish Local Government, Scotland*

ABSTRACT

Prior to the establishment of the Knowledge Management (KM) strategy, the British Council defined knowledge as objects. Knowledge sharing was about sharing documents and information on the intranet or via global databases. Since December 2002, Dervin’s Sense-Making Methodology has been applied to manage knowledge. Knowledge is seen not as a product that can be transferred from one colleague to another but as a communication practice. This means that shared knowledge has to be interpreted and made sense of by its recipients through genuine dialogue. During this phase of KM implementation, the focus shifted to linking up colleagues and providing space for dialogue through building global communities of practice and virtual teams. This article presents an example of how we have used the theory of Social Networking Analysis as a diagnostic tool to promote knowledge sharing among our newly formed 30-people global leadership team. The three steps we have taken to carry out the exercise and its limitations also are discussed.

BACKGROUND

The purpose of the British Council is to build mutually beneficial relationships among people in the UK and other countries and to increase appreciation of the UK’s creative ideas and achievements. Much focus has been on sharing knowledge and experience with customers. In order to take the organization to another level, the British Council promotes knowledge sharing among its 7,000 employees, who are located in 109 countries. The ultimate aim is to empower staff to get the knowledge they need to serve their customers to the highest standard possible.

The Knowledge Management (KM) program was launched officially in December 2002 with the appointment of the new Director of Knowledge Management. Following a comprehensive six-
month knowledge audit exercise, the global KM strategy was approved by the Senior Management Team in December 2003. The KM vision was to enable the British Council to develop and deliver world-class products and services to its customers by effectively sharing and utilizing collective knowledge. This will be achieved by finding the best ways to connect its employees with each other and by providing them with easy access to relevant documents and resources.

Over the last two years, we have launched a number of projects to increase awareness of KM and to get the buy-in of senior management for the program in order to invest in the tools and approaches needed to improve global knowledge sharing. Specific KM projects that are beginning to embedded into the organization include:

1. Knowledge audit conducted using Dervin’s Sense-Making Methodology (Dervin, 1992; Cheuk & Dervin, 1999)
2. Development of Knowledge Management strategies for business units
3. Building communities of practices using seven-phase methodology (Cheuk, 2004a, 2004b)
4. Enhancement of the intranet, collaboration tools, and global databases
5. Applying Social Networking Analysis to support collaborative working (Anklam, 2003; Cross & Parker, 2004)
6. Applying narrative techniques to conduct project debriefs

In 2005, KM was recognized widely as an enabler to deliver the British Council’s overall business strategy. More than 100 knowledge champions worldwide have attended training on KM, and more than 70 global communities of practice have been developed.

**BUSINESS CONTEXT**

During 2004-2005, the overseas operations of the British Council were significantly restructured. Thirteen regions have been introduced to replace the existing 109 country operations that each were managed as individual entities. Each new region is made up of a number of existing country operations.

Thirteen regional directors were appointed. They have to work closely with the 17 senior management team members based in the UK in order to set strategic direction for the organization. This 30-person team is referred to as Global Leadership Team (GLT).

The restructuring provides an excellent opportunity to promote knowledge sharing beyond country operations as well as to promote knowledge sharing between overseas operations and the UK headquarters. However, it also presents a challenge. Any organizational restructure leads to the creation of new teams, which can be to the detriment of any existing knowledge-sharing culture. This presents a challenge to the KM team.

**WHAT IS THE NATURE OF KNOWLEDGE?**

A review of the KM literature largely defines knowledge using Nonaka and Takeuchi’s (1995) definition of tacit knowledge (i.e., knowledge in a person’s head, which has a personal quality and is hard to formalize and communicate) and explicit knowledge (i.e., knowledge that is transmittable in formal, systematic language). Sutton (2001) defines the latter as codified knowledge (i.e., knowledge that can be written down) and uses it interchangeably with information. Taken this view, KM systems were created to capture the knowledge of experts. The capture approach continued with an emphasis on capturing knowl-
Using Social Networking Analysis to Facilitate Knowledge Sharing in the British Council

data. Hildreth & Kimble, 2002).

Wilson (2002) argues that one should not use the terms  
information and knowledge inter-
changeably. He proposes that everything outside 
the mind that can be manipulated in any way is 
defined as data or information. They take the 
form of papers in a journal, e-mail messages, 
manuscript letters in an archive, and so forth. 
However, knowledge (i.e., what we know) never 
can be managed, except by the individual knower 
and, even then, only imperfectly. Wilson (2002) 
defines knowledge as what we know. He elaborates 
that knowledge involves the mental processes of 
comprehension, understanding and learning that 
go on in the mind and only in the mind, however 
much they involve interaction with the world 
outside the mind, and interaction with others. 
Whenever we wish to express what we know, we 
only can do so by uttering messages of one kind 
or another—oral, written, graphic, gestural, or 
even through body language. Such messages do 
not carry knowledge; they constitute information, 
which a knowing mind may assimilate, under-
stand, comprehend, and incorporate into its own 
knowledge structures. These structures are not 
identical for the person uttering the message and 
the receiver, because each person’s knowledge 
structures are, as Schutz (1967) puts it, biographi-
cally determined. Therefore, the knowledge built 
from the messages never can be exactly the same 
as the knowledge base from which the messages 
were uttered. He argues that knowledge cannot 
be managed and that Knowledge Management is 
a nonsense concept.

Dervin (1992) takes on an alternative view that 
it is not important to distinguish information and 
knowledge from a communication perspective, 
because knowledge is not an object but rather 
a communication process or flow. Knowledge 
(regardless of what label you give it) is anything 
that makes sense to the users through a dialogic 
communication process. Knowledge can be de-

defined only from the users’ perspectives (Dervin, 
Foreman-Wernet, & Lauterbach, 2003).

Prior to the establishment of the KM strategy 
in December 2003, the British Council took the 
traditional view and defined knowledge as in-
formation and documents that can be managed 
as objects. KM was seen as sharing documents 
and information on the intranet or via global 
databases.

The launch of the KM strategy in December 
2003 employed Dervin’s (1992) Sense-Mak-
ing theory to provide the organization with an 
alternative perspective in knowledge sharing. 
Knowledge is no longer seen as a product that 
can be transferred from one colleague to another 
but as a two-way communication practice. In ad-
dition to connecting employees to information 
using KM systems, we begin to focus on link-
ing employees with employees, in particular to 
facilitate genuine dialogue between employees. 
Instead of asking the question “what information 
should we manage?” we begin to ask the question 
“who should be linked up?” in order to maximize 
business outcomes.

During this phase of KM implementation, 
the focus was on providing space for dialogue 
through building global communities of practice 
and virtual teams to deliver strategic programs. At 
the British Council, we put in facilitators to man-
ge these communities and supported them with 
Web-based collaboration tools. Social Networking 
Analysis also was introduced as a diagnostic tool 
to support team building as well as to evaluate 
the performance of the communities.

**THEORY BEHIND SOCIAL NETWORKING ANALYSIS**

Social Network Analysis has emerged as a set 
of methods for the analysis of social structures, 
methods that are geared specifically to an investi-
Using Social Networking Analysis to Facilitate Knowledge Sharing in the British Council

gation of the relational aspects of these structures. The use of these methods, therefore, depends on the availability of relational rather than attribute data (Scott, 1992).

Social Networking Analysis can be used to address various organizational issues such as supporting partnerships and alliances; assessing strategy execution; improving strategic decision making in top leadership networks; integrating networks across core processes; promoting innovation; ensuring integration post-merger or large-scale change; and developing communities of practice, personal networks, and leadership development (Cross, 2004).

The importance of social networks is highlighted by Cross, Parker, Prusak, and Borgatti (2001), who found that despite easy access to a world-class KM system, 85% of managers got information (that had an impact on the success of a project) from their personal networks.

Snowden (2002) emphasizes that knowledge is stored not just in documents. He argues that “we always know more than we can say, and we can always say more than we can write down” (p. 104). He suggests that we must pay attention to managing social networks—this is how knowledge in people’s heads is transferred naturally and rapidly across an organization.

Informed by these theories and associated research findings, the British Council has introduced Social Networking Analysis as a diagnostic tool to improve strategic decision making in the newly-formed Global Leadership Team.

WHAT HAVE WE DONE TO PROMOTE KNOWLEDGE SHARING?

We started with 13 newly appointed regional directors in charge of setting strategic direction for the British Council and the delivery of products and services overseas. They are geographically dispersed, they have to work closely with the 17 senior management team members in the UK headquarters, and this Global Leadership Team will meet only twice a year in the UK.

The KM team wanted to help them to understand what knowledge sharing across countries and regions was really like. They needed to understand what knowledge had to be shared, how to help them and their staff get their work done, and what needed to be in place in order to make knowledge transfer effective. The KM team wanted to give them practical experience before they went on to introduce new knowledge-sharing approaches to the teams within their region.

A series of activities were designed to promote knowledge sharing among the 13 global leaders as well as between them and the 17 senior management team members in the UK. They included:

1. An initial community-building meeting to help them get to know one another.
2. An audit exercise to find out what knowledge, resources, expertise, and help they need to get work done.
3. The establishment of a Web-based collaboration site.
4. The appointment of a community facilitator.
5. The establishment of an events calendar.
6. Carrying out a Social Networking Analysis exercise to identify opportunities and gaps.

HOW DID WE CONDUCT THE SOCIAL NETWORKING EXERCISE?

The KM team conducted a Social Network Analysis exercise for the 30 global leaders as part of a global leadership development event. The aim was to help the team to visualize its existing relationships and to allow the group to reflect on how the group members network with one another.

There are three steps to complete this exercise:
Using Social Networking Analysis to Facilitate Knowledge Sharing in the British Council

• **Step 1.** A data collection template was developed and circulated to the 13 regional directors (as well as the 17 senior management team members in the UK). They were asked to complete the form prior to the event (Appendix 1).

• **Step 2.** The findings were analyzed and presented to the group during the 60-minute Knowledge Management session.

• **Step 3.** The participants were given the opportunity to discuss the findings and to come up with interventions to focus on during the next three months (Appendix 2).

The KM team emphasized that Social Networking Analysis is best employed as a diagnostic tool to generate discussion relating to team building and communications. It is not meant to evaluate the performance of the group or individual members. A step-by-step guide to the process is presented next.

**Step 1. Data Collection**

The template shown in Figure 1 was circulated to the 13 leaders (and to the 17 senior management team members in the UK) before and during the

---

**Figure 1.**

<table>
<thead>
<tr>
<th>Name</th>
<th>Ref No</th>
<th>Scoring</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 = Never</td>
<td>1 = Once in two months</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alan</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andrew</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>James</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martin</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Michael</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morna</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philip</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rob</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rod</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosemary</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sarah</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stephan</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reflect on your interaction with the following global leaders since September 2004 and answer the following questions

1. **Q1:** To whom and from whom do you send and receive information, documents, plans, and other resources?
2. **Q2:** With whom do you have informal discussions about your work and/or new ideas?
event. Altogether, 30 colleagues were given the form. The template was designed to be simple and self-explanatory. Only two questions were asked in this case:

**Q1.** To whom and from whom do you send and receive information, documents, plans, and other resources?

**Q2.** With whom do you have informal discussions about your work and/or new ideas?

**Step 2. Data Analysis and Visualization**

Only 23 colleagues out of 30 global leaders completed the data collection template. The Knowledge Management team used UCINET software to visualize the data using Social Networking Analysis (SNA) Maps. During the event, the global leaders first were given an introduction to Social Network Analysis (SNA) and then presented with the findings in terms of three anonymous daily, weekly, and monthly SNA maps. Each map demonstrated the frequency of colleagues’ contacts with one another for formal information exchange. The nodes representing colleagues based in the UK and overseas are light grey and dark grey respectively. Figure 2 shows one of the SNA maps.

**Step 3. Discussion**

The leaders were invited to discuss the following questions:

*Figure 2.*
Using Social Networking Analysis to Facilitate Knowledge Sharing in the British Council

1. What patterns do you see?
2. Where do you think you sit in the SNA map?
3. What do you see as the key strength of this network?
4. What do you see as the potential weakness of this network?

We respected the privacy of the data, as it disclosed the relationship between colleagues. At first, we showed a diagram that did not provide names of individual nodes. To our surprise, all the participants asked for the results to be disclosed during the event and immediately gave consent for their names to appear on the SNA map.

As a result, we presented the map shown in Figure 2. They were given time to reflect on the above questions again and then to discuss these questions:

1. As a group, what needs to be changed in three months in order to achieve the Global Leadership Team’s objectives?
2. As an individual, what would you like to change in three months?

WHAT DID THE GROUP LEARN?

The newly appointed global leaders agreed that the SNA map represents a reasonably accurate reflection of the situation at that time, given that the majority of global leaders were newly in post. In addition, they highlighted the following issues regarding knowledge flow:

- There were strong relationships between UK-based staff.
- There was relatively little overseas/UK interface.
- Only a few overseas leaders were talking to one another.
- There was little difference in terms of formal and informal networking patterns.

They also reflected on the strength and weaknesses of their network. The strength was that the monthly SNA map showed the volume of communication that already was taking place. The weaknesses were:

- The preponderance of nodes in the headquarters. This might hinder widespread communication of messages.
- More networking was needed between the global leaders.
- There was a need to set clear, defined tasks to ensure that communication takes place.

As a result of the discussion, they identified some actions to improve networking:

- Several leaders agreed to form a mini group to work together on issues.
- A monthly Web meeting was arranged to allow the group to discuss issues.
- The need to nurture existing subgroups (e.g., several regional leaders were already discussing issues informally with one another). They wanted to make an effort to keep that going.

WHAT IMPACT HAS RESULTED?

As a result of the interventions that the global leaders identified to improve networking among themselves, on top of their busy schedule they made a conscious effort to touch base with one another through online knowledge-sharing sessions. Over time, they have built a better understanding of one another and have shared the challenges they face and how they overcame them in their region. Improved networking among the Global Leadership Team members also has led to unexpected outcomes, whereby knowledge exchange is not limited to topics on the set agenda but takes place on a more informal basis covering a range of other topics.
In summary, through improved social networking, the global leaders have led by example and have contributed to improve knowledge sharing in a number of ways:

1. It has increased the number of documents shared on the collaborative Web site (as a result of the need to exchange documents to prepare for or as a follow-up to a networking event).
2. The global leaders share important projects they are implementing in their regions and which employees are leading them. This information, in turn, is shared by the regional director with their regional team, which has resulted in increasing networking between managers in different regions.
3. It has opened up the eyes of global leaders to the power of social networking through good facilitation. Many regional directors have expressed an interest in conducting a similar social networking exercise with their own regional management team.

LIMITATIONS

A number of participants pointed out the following limitations of the social networking map:

- A successful SNA map requires a 100% response rate. All participants must fill in the data collection sheets. This is difficult to achieve. At the British Council, we achieved this through distributing the form at a compulsory event. However, only 23 out of 30 Global Leadership Team members completed the data collection template.
- There were many nodes on the network (especially in the UK). This can lead to the false conclusion that the more nodes there are, the stronger the network is.
- A number of leaders mentioned that the SNA map only reported the situation at that time. The maps should be used with care, as they only can provide a snapshot at any one point in time. The group thought it would be useful to go through the same exercise again at a later date in order to see the shifts in type and strength of relationships.

The other limitation is that in this social networking exercise, only 23 people participated. When applied to a larger number of staff (e.g., more than 1,000), the resulting SNA map could be more complex, and the discussion as to what a person can do as an individual to change and improve networking could be more difficult. Additional applied research has to be done to understand the value of SNA to business.

CONCLUSION

This article presents an example of how the British Council has used the theory of Social Networking Analysis as a practical tool to support our Knowledge Management program. It proves that SNA exercises are simple to carry out and that the results can provide a focal point for discussion in improving knowledge flow.

The global leaders who attended the session agreed that it was worth completing the exercise and that the SNA maps provided them with alternative perspectives on their own knowledge flow and networking habits. They began to recognize the need to balance the sharing of knowledge through documents against people-to-people networking. It helped them to improve their understanding that Knowledge Management as a subject encompasses more than document exchange. They recommended that SNA exercises be adopted for supporting team building at a regional level.
REFERENCES


ENDNOTES

* The author was Director of Knowledge Management, British Council when this article was written

1 One community of practice qualified as a finalist in the KM category of the Information Management 2004 Award. Another
community of practice received a commendable award in the KM category in the Information Management 2005 Award.

Chapter 5.17
Collaboration Challenges in Community Telecommunication Networks

Sylvie Albert
Laurentian University, Canada

Rolland LeBrasseur
Laurentian University, Canada

ABSTRACT
This article reviews the literature on networks and, more specifically, on the development of community telecommunication networks. It strives to understand the collaboration needed for innovative projects such as intelligent networks. Guided by a change management framework, collaboration within a community network is explored in terms of the formation and performance phases of its development. The context, content, and process of each phase is analyzed, as well as the interaction of the two phases. User involvement and technology appropriation are discussed. Collaboration challenges are identified and linked to the sustainability of the community network. Policy makers are presented with a model that gives some insight into planning and managing a community network over time.

INTRODUCTION
Collaboration in networks and managing performance across organizations has gained the attention of researchers (Huxham & Vangen, 2000). Our comprehension of collaborative networks has progressed substantially over a couple of decades (Oliver & Ebers, 1998), but it lacks integration (Ebers, 2002).

Collaborative networks cover a range of purposes such as innovation requiring heavy investment in R&D, international ventures, and the delivery of public services like health and education. This article is focused on telecommunication networks that operate within a physical and shared community space. The more ambitious community networks aim to become “intelligent” communities with broad participation and significant impact on the local social and
economic development. To understand them as a
dynamic phenomenon, a framework is needed that
can accommodate and organize the conceptual
pillars of organizational environment, structure,
culture, leadership, and management. Pettigrew
(1992, 1987) offers such a framework, and Ebers
(2002) and LeBrasseur et al. (2002) demonstrate
its effective application.

Organizations in all sectors have become more
interested in inter-organizational collaboration
to encourage synergy, innovation, and economic
development. Although there are many pockets
of successful collaborative efforts, there is a
continuing need to identify the challenges and
opportunities inherent to community networks.
With this focus, this article is divided into four
main sections. First, collaborative networks are
defined and described, and community telecom-
munication networks and their potential for sup-
porting intelligent communities are analyzed.
Second, key collaboration challenges that impact
on the development of a community network are
introduced. Third, the literature is reviewed and
organized according to the context, content, and
process involved in these community networks
during their two phases of development—formation
and performance. The collaboration challenges
present in each phase of development are explored,
including challenges that the users experience.
Fourth, the article concludes with policy implica-
tions for network planners.

TELECOMMUNICATION NETWORKS
AS AN EXAMPLE OF COLLABORATION

Collaboration is the pooling of resources (e.g.,
information, money, labour), by two or more
stakeholders or partners, to solve a set of prob-
lems, which neither can solve individually (Gray,
1985). It involves an interactive process whereby
organizations, using shared rules, norms, and
structures, act or decide on issues related to a
problem domain (Wood & Wood, 1991). The
intentional goal-oriented collaborative arrange-
ment that emerges is that of a network (Poyhonen
& Smedlund, 2004).

Networking represents a particular form of
organizing or governing exchange relationships
among organizations and is an alternative to
Network partners maintain their autonomy and
retain residual property rights over their resources
that have been pooled to achieve mutually agreed
outcomes (Bailey & McNally-Koney, 1996; Brown
et al., 1998; Gray & Hay, 1986; Huxham & Van-
gen, 2000; Oliver & Ebers, 1998). The principal
coordination mechanisms for allocating resources
are negotiation and concurrence. Informal social
systems, rather than bureaucratic ones, coordinate
complex products or services and reduce uncer-
tainty (Jarillo, 1988; Jones et al., 1997).

Networks have gained in importance over the
last two decades. For the private sector, global-
ization and the speed of change have encouraged
collaborative efforts. For government, down-
loading since the 1990s has forced new ways to
view management of programs and services for
resource maximization (Bradford, 2003; Bailey
et al., 1996). Municipalities and regions have also
demonstrated an increased interest in collabora-
tion efforts and network development to attract
new opportunities and maintain their competitive
advantage. Collaborative networks typically in-
crease the scale and visibility of program efforts,
increase support for projects, and leverage capital
to enhance feasibility, speed, and effectiveness
(O’Toole, 1997). Synergy is achieved through
improved resource management and intensive
exchanges on specific projects.

To achieve synergistic gains and programming
enhancements from sharing resources, risks, and
rewards, stakeholders need to shift their focus
toward collaborative rather than competitive
advantage (Lowndes & Skelcher, 1998). Too
often in the past, public sector organizations
built independent silos and their private sector
counterparts viewed potential partners as competitors rather than collaborators. Public policies dealing with ambitious or complex issues, like community sustainability, are likely to require networked structures that allow for the pooling and mobilization of resources from both private and public sectors within a government policy initiative (O’Toole, 1997).

Community telecommunication networks reflect the trend in western society away from bureaucratic government to network governance (Sorensen, 2002): the latter delivers more services efficiently with less risk and uncertainty (Considine & Lewis, 2003; Jones et al., 1997). Stakeholders and collaborators include municipalities, health, education, social services organizations, and private sector organizations. These networks are part of a wider agenda to increase the country’s capability for the knowledge-based economy.

There are several kinds of community networks (Gurstein, 2000; Pigg 2001), ranging from those serving a restricted membership (usually called private networks) to those serving a broader segment of the community or region. A private network may, for example, link several schools and/or municipal sites, and members would include the students, administration, and staff of these organizations. In contrast, a community network is built on a platform that gives broad access to community citizens, businesses, and agencies; it encourages many stakeholders to become a user and service provider. These stakeholders may come together simply to upgrade an aging infrastructure, especially when market forces cannot be relied upon to meet community needs or to co-build economic foundations. Leading communities strive to build partnerships and synergy to overcome barriers to access, job creation, and innovation (Agres et al.,1998; Eger, 2001; Tan, 1999; Industry Canada, 2002a, 2002b). Community networks facilitate information dissemination, discussion, and joint activity by connecting neighbours, creating new opportunities, and empowering residents, institutions, and regions (Carroll & Rosson, 2001; Igbaria et al., 1999; Canadian National Broadband Task Force, 2001).

A community network has four basic components: a telecommunication infrastructure with broadband capability, applications or content, devices (such as computers, cellular telephones, i-pods, and blackberries), and users. The development of a community telecommunication network typically occurs through a governing board representing the needs of users, which is supported by a small management structure (e.g., executive committee and network manager). The network relies on information and communication technologies (ICTs) and allows the community to import and export knowledge, encourage innovation, and overcome distance. The opportunities for economic and social development are contingent on attracting many users and creating a culture of “digital” use. The network must fulfill user needs and be attentive to their requirements, which may include a fair price, access to computers and the Internet, and training and education.

Infrastructure investment in the telecommunication network aims for the widest possible coverage of the community and region, with the constraint of reasonable cost. Investment also tries to ensure that users have access devices; some users have modest means, and schools and other organizations may have budget constraints. On the human resources front, technical training of local staff may be required to install the infrastructure and devices, and provide support to users. Organizations may need to re-design processes in order to meet the changing needs of their supplier and distribution partners, and to offer services online to end-users. The transformation effort may also require promotion campaigns to attract both individual and organizational users. These many resource challenges imposed on the community require a collaborative effort to pool resources and find innovative solutions.

A community network has users at the individual, organizational, and community levels of
human activity and endeavours. Individuals or end-users use the network to communicate with friends, play games, access information, obtain training, and occasionally innovate. Organizations are often stakeholders and use the network for a wide variety of purposes (Waits, 2000), including providing information and services, and selling online. They are intermediate users (Williams et al., 2005) and are the drivers of the development of the network. These organizations are the channels through which collective innovation is exercised and community change takes place (de la Mothe, 2004; Rycroft, 2003). At the community level, these individuals and organizations create aggregate demand and use of the network, and determine the sustainability of the network. The challenge is to create a culture of “digital” use that is integrated into the broader culture that is shared by community members.

During the development of the network, user involvement can be traced through participation in articulating a “digital” vision for the community, in the purchase of access devices and services that will connect users to the network (e.g., telephone, cable, wireless, computers, and Internet), and in the utilization of applications being made available through these access devices. Users may also be involved in creating employment by innovating on the network configuration, and in helping to create a culture of use by providing additional attractive applications.

Good governance requires legitimacy with an appropriate range of stakeholders, and involves building cohesion and commitment. Relationships are voluntary, and network survival depends upon the collective will and commitment of the stakeholders. The intentionally-planned network takes on a collaborative structure composed of local residents, non-governmental organizations, private sector businesses, and government. The stakeholders create a product that reflects the concerns, priorities, and aspirations of the local population. If the infrastructure, devices, and applications meet the needs of the users, a culture of “digital” use emerges as an organic extension of existing community ways and practices. Without broad participation, the network is likely to reflect narrow interests and weaken the community’s social sub-system, which in turn will limit the economic success of the network.

A sustainable community telecommunication network makes consistent and positive contributions to the economic and social development of the community (ITU, 2003), thereby enhancing the community’s capital base. In large measure, these positive outcomes depend upon the collaboration of partners. They also reinforce the efforts invested in collaboration. Networking allows individuals, professionals, and entrepreneurs to access information and knowledge, learn about a wide range of issues, recognize opportunities, and achieve innovative products and services (Suire, 2004; Martin & Matlay, 2003; Corbett, 2002; Ardichvili & Cardozo, 2000; Kickul & Gundry, 2000). Whether a community network realizes its potential depends upon how well it is developed.

The above discussion portrays the formal structure of a community network as a fluid organization composed of volunteers with the purpose of facilitating the community’s transition and participation in the information society. However tempting, this viewpoint is non-critical in nature; it ignores the community context and processes by which the network emerges (Pigg, 2001; Day 2002).

**COLLABORATION CHALLENGES FOR COMMUNITY NETWORKS**

Communities around the world have demonstrated that transformation is possible using network technology. For example, Sunderland (UK) reduced unemployment from 30% to 4% by moving from a shipbuilding and coal industrial base to a knowledge and technology economy. Similarly, Spokane Washington (USA), once a railroad town reliant on natural resources, dra-
matically improved the fortunes of its downtown by installing the highest density of broadband in the country. In Tianjin (China), a major push on broadband connectivity was accompanied by rapid user growth, from 20,000 to 2,700,000 in two years. Their stories make ample reference to the intensive collaboration of many actors, but the patterns of influence are not well articulated. Bell (2001) compared six urban communities noted for their telecommunication achievements and identified two effective patterns of collaboration: (1) a comprehensive and formal plan, and (2) a coherent pattern of individual initiatives. Similarly, Williams et al. (2005) reviewed numerous ICT initiatives, both small and large, and emphasized the overlapping nature of the planning, implementation, and use stages of development. These patterns are explored under the phases of network development section of this article.

Individuals and organizations involved in the creation of a community network face four collaboration challenges:

1. Defining and agreeing on the transformation effort (includes vision, transformation, and planning)
2. Assembling and mobilizing resources (includes interdependence, tasks, and structure)
3. Assembling and mobilizing trust (includes prior experience, communication, and distribution of power among collaborators)
4. Balancing leadership and collaborative management (includes the broadening requirements of the network, user appropriation, and power).

These challenges are tied to the coherence and adaptability of the network, and specifically to the dynamic relationship between the formation and performance phases of its development. Collaboration is inter-woven in each of these challenges. Network sustainability is achieved by collaboration efforts that evolve during the network’s development.

**PHASES OF DEVELOPMENT OF A COMMUNITY NETWORK**

We propose that network development takes place in two phases that are iterative in nature. Phase 1, the formation of the community network, is marked by the emergence of a leader and/or a board of directors, to respond to environmental pressures. These pressures may occur as a result of globalization and the need to remain competitive in the face of other communities or regions. It may occur as a result of downsizing or social development pressures (e.g., lack of medical practitioners, youth out-migration). The broad goals of the network are developed, including a representation of the future user. Phase 2, network performance, involves the concrete objectives and steps that the board takes to achieve the community goals that were agreed upon and the measures taken to attract and retain users. User involvement can and should take place in both phases of development.

Smaller communities need collaborators to solve a wide variety of challenges including infrastructure availability. Larger communities tend to have more resources and thus need collaboration to resolve economic and social pressures rather than infrastructure issues. In this second phase, the network can develop a culture and structure that gives meaning and coherence to a variety of projects. Some communities are more liberal and hands-off, allowing the private sector and citizens to develop content and opportunity. Others intentionally plan a vision of community transformation based on an improved telecommunications infrastructure. Phase 1 depends highly on leadership dynamics whereas Phase 2 is closer to managerial dynamics but with a distinctive collaborative flavor. These two phases are interdependent over time in that formation sets the stage for performance, and performance impacts on the board and leadership dynamics. Positive outcomes at the performance phase consolidate the dynamics of the formation phase; negative
outcomes challenge the board and leadership and initiate a re-formation phase. This iterative process was demonstrated in the feedback loop identified by Arino and de la Torre (1998) and Thomas (1993).

Because networks are fluid in nature (pooling from existing resources, changing membership, and varied timelines) and focused on both results and relationships, two interactive phases are considered sufficient. The two phases are supported by case studies of strategic alliances (Doz, 1996) that found that successful partners actively exchanged information, re-evaluated the project (in terms of efficiency, equity and adaptability), and consequently readjusted the initial conditions of their cooperation. They are also consistent with the ICT social learning findings of Williams et al. (2005).

**Formation Phase**

The push and pull factors in the environment impact on the community members and prompt them to consider uniting their forces to address the issue or opportunity that has been identified. Under the leadership of a visionary, and through ample interpersonal communication, a group is assembled that represents the initial membership of a potential network. If a consensus on vision and goals is attained, the group becomes the founding board of a network and plans for the performance phase. The principal outcome is a collaborative base on which to build the network. Table 1 provides an overview of the critical factors present in the formation phase.

**Context of Formation**

The outer context or environment includes factors such as economic, political, culture, demographics, funding opportunities, pressures from government agencies, and technology innovation trends (Agres et al., 1998; Bailey & McNally-Koney, 1996; Igbaria et al., 1999; Keenan & Trotter, 1999; and Tan, 1999). Global competitiveness and turbulence are the underlying catalysts for creating networks for organizations, communities, and governments (Poyhonen & Smedlund, 2004; Scheel, 2002).

Interdependencies exist because organizations possess or control vital resources (material, human, political, structural, or symbolic) and thus are the source of environmental pressures for one another (Wood & Wood, 1991). Organizations seek to reduce these pressures and manage

---

**Table 1. Influencing factors at the formation phase**

<table>
<thead>
<tr>
<th>Context</th>
<th>Content</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Economy</td>
<td>• Vision</td>
<td>• Values</td>
</tr>
<tr>
<td>• Social/cultural</td>
<td>• Power</td>
<td>• Expectations</td>
</tr>
<tr>
<td>• Political</td>
<td>• Board Membership</td>
<td>• Goals</td>
</tr>
<tr>
<td>• Urbanization</td>
<td>• Concept of Sustainability</td>
<td>• Planning</td>
</tr>
<tr>
<td>• Funding</td>
<td>• User representation</td>
<td>• Leadership</td>
</tr>
<tr>
<td>• Technology</td>
<td></td>
<td>(transformational, visionary)</td>
</tr>
</tbody>
</table>
Collaboration Challenges in Community Telecommunication Networks

the interdependencies by gaining control over crucial resource supplies. The network form, as opposed to markets and hierarchies (e.g., vertical integration), provides a neutral space within which organizations can meet to explore solutions and synergies.

International bodies such as the World Bank (1999), the United Nations (1998), and OECD (1997) have adopted the paradigm of the information society as a guide to many of their development policies. According to Castells (1996, 1997, 1998), ICTs have produced a network society in which mobilizing knowledge and information have become more important than mobilizing physical resources. He argued that both organizations and individuals can benefit from electronic networks; they support the development and dissemination of knowledge and information, and facilitate innovation. Castells warns that these changes are accompanied by growing wealth disparities, social fragmentation, and dislocation. Governments are addressing these concerns, in part, by financially supporting the creation of community networks with broad accessibility. Locally, these new opportunities are often communicated through the chamber of commerce and other economic development agencies to mobilize or inspire stakeholders into action.

Communities come in all sizes and density, and all are influenced by the urbanization trend. Rural settings are witnessing the exodus of their youth and an erosion of their economic base as cities attract both talent and investment, including initiatives in telecommunications (OECD, 2004). Recent studies of Canadian rural communities concluded that ICTs can act as enablers of community building and development processes (New Economy Development Group Inc., 2001; Canadian Advisory Committee on Rural Issues, 2004). Given that the Internet is content-rich, offers online education, facilitates social networking, and offers a platform for the creation of new enterprises and the expansion of existing ones, the viability of the digital community network becomes crucial for the future of small communities.

When governments create generous programs to create community networks (e.g., Brown et al., 1998), communities are pressured to apply for capital funds even when they may not have the organizational and resource capacity to sustain the network. Smaller communities have relatively fewer and less diverse resources and a push-style policy may be the only way to spur action. Another example of a push factor is when a major telecom firm seeks a community partner for a demonstration project, or when the private sector chooses to make an investment to upgrade its infrastructure. The telecom supplies the ICTs, but the community stakeholders still need to demonstrate and plan on how the technology can be applied to personal and organizational purposes. Often, infrastructure is built and languishes until there are other pressures in the environment of the community, such as closure of a major employer or the arrival of a strong champion. At other times, communities struggle with the lack of open access that inhibits economic development and competition. Pushing for open access can discourage the involvement of incumbent carriers, at least at the onset. The key here is to evaluate how context issues can stimulate communities into action toward their transformation.

Content of Formation

Stakeholders need to find the community vision attractive and see a benefit for themselves and for their organization. When the problem is broad in scope and complex, such as economic development, it requires a larger variety of stakeholders with legitimate interest to devise solutions and bring sufficient resources to bear. Stakeholders must have the right and the capacity to participate, and include organizations with legitimate power as well as those who will be affected by the network.
Collaborative action necessarily involves interdependence between individuals and organizations (Ouchi, 1980) and can yield both intangible (e.g., image of citizenship) and tangible benefits (e.g., cost reductions and additional revenues). Interdependence is strongly linked to the vision of the network and the factors motivating stakeholders. It allows for an exchange among stakeholders that is built on trust, and an understanding of mutual benefit or advantage. According to Okl and Young (1997), the more ties an organization has to others in a network, the less likely is it to act opportunistically. Blois (1990) argued that collaborators should engage in bargaining on who will accept responsibility for certain elements of the transaction costs. They must come to the table understanding their role and develop a level of interdependence and mutual benefit in order to sustain the network effort.

The economic and social exchanges that take place are mediated by mutual trust. Ring (2002) distinguishes between “fragile” and “resilient” trust. The former is typical of opportunistic settings such as markets and involves the formal processes of negotiation, transaction, and administration. In contrast, the latter is the foundation of successful networks and is based on the informal processes of sense-making, understanding, and commitment. However, prescribing resilient trust does not ensure that it takes place. Ring proposed that it will emerge when the participants have a shared history of experience and when reputations for reliability are well established. On the other hand, Doz (1996) has documented the role of trusted intermediaries in helping other participants to shift gradually from fragile to resilient trust. We conclude that if a community has rich social relations, it can establish resilient trust early, but that parachuting in partners and stakeholders makes fragile trust more likely. However, if trusted intermediaries become involved, they can build the level of trust within the network.

There is a need for legitimate authority, credibility, and multiple memberships (Bailey & McNally-Koney, 1996; Gray & Hay, 1986) if a sustained transformation is to occur. Jones et al. (1997) have argued that networks need to restrict membership access and choose its members according to their reputation and status. Important stakeholders may choose to join a network or a project because of the presence of other members. A smaller number of leaders may allow the network to realize quick wins, reduce coordination costs, and improve interaction frequency. One could argue that success will breed success—trust will increase, motivation will increase, and faster output can be generated. This view is less applicable to community networks where innovation, legitimacy, and broad reach is critical and leads to a large membership and numerous exchanges. Therefore, a smaller, more restricted network may be mobilized quickly and act efficiently, but be less effective in producing varied output. The larger network may slow the pace of change, but may be important enough to attract accomplished leaders. Structure issues become important in managing a larger group of stakeholders and are discussed in the performance phase of the network.

Another content issue is sustainability. Stakeholders want to know, “How much will it cost,” but few ask “How will the network become sustainable in the long-run?” Sustainability is a function of revenues (stemming from the use of the infrastructure and its applications), and the costs of the network (human resources, equipment, and materials). There are opportunities for synergistic gains when partners chose to purchase as a group, or share the operating costs. At the formation phase, the concept of sustainability is hazy, but becomes clearer as projects develop during the performance phase. Nevertheless, the board must carefully address the sustainability issue early to ensure that it becomes incorporated into their common frame of reference.

In the formation stage, the planning includes an explicit model of future users, their communication needs, and their likely use of the telecommunication network. Williams et al. (2005, p. 112,
Figure 5.2) identify ways for direct involvement of users, such as user panels, market research, and trials. They also identify sources of indirect evidence about users through information on demand and markets for similar products, and competitive offerings. With the additional input of board members who understand their community, a representation of the users is developed. This user-centered approach is helpful in guiding the design of the system and identifying training and promotion requirements. However, Williams et al., emphasize its limitations and the design fallacy that it breeds: “the presumption that the primary solution to meeting user needs is to build ever more extensive knowledge about the specific context and purposes of an increasing number and variety of users in the technology design” (p.102). The idea of perfect user representation ignores the reality that users are active agents and appropriate the technology later, primarily in the performance phase of network development.

Communities would be wise to involve users in all facets of their formation stage, but users are often thought of as passive participants that can be surveyed for the eventual purchase of devices or services at the performance stage. Yet, users have concerns over ownership, access and distribution of information, privacy, security, and copyrights (Agres et al., 1998), and most of these issues need consideration early on. However, the design fallacy mentioned above emphasizes the limitations of comprehensive user involvement in the formation phase.

**Process of Formation**

Leaders and champions can enhance or constrain the development of a community network (Industry Canada, 2002; Jones et al., 1997; Huxham & Vangen, 2000). Leaders tap into the collective awareness of the community stakeholders and mobilize the initial change efforts by supplying a vision and practical steps to realize it (Bailey & McNally-Koney, 1996; Roberts & Bradley, 1991). Sustaining collaboration depends on the emergence of a common view of the community and shared organizational values. Leaders and champions play a role in consolidating and expanding the collaborative initiatives, but a wider involvement is needed to foster innovation. It is important to have a community cross-section of members as well as individuals with sufficient power to rally other stakeholders. The parties must freely participate, knowing and agreeing on who is involved and in what capacity (Glatter, 2004; Roberts & Bradley, 1991); prior experience and trust facilitate the membership drive.

Network goals are created, implemented, and modified through purposeful social construction among network stakeholders and partners (Van de Ven & Poole, 1995; Ring & Van de Ven, 1994). Network effectiveness may be defined as the harmonization, pursuit, and attainment of the goals sought by the various stakeholders and partners. With diverse stakeholders, it becomes difficult to satisfy all parties equally; therefore, managing expectations and potential conflicts help to maintain the social cohesion of the network. Members will likely persist so long as they can positively identify with the intermediate and long term outcomes, whether they are social or economic in nature.

According to Hardy and Phillips (1998), when individuals come to share a vision of the issues and the solutions, they become stakeholders and begin to create a collective identity with mutually agreed upon directions and boundaries that, in time, may become a permanent network. The catalyst is a transformational leader who encourages collaboration as a means to create synergy for innovation, growth, or to protect against future turbulence. Engaging the stakeholders in a planning exercise can address their many concerns; tasks and roles can be organized and assigned within the network to fit their expectations. Because work is complex and elaborate in networks, planning and coordinating task-specialized activities is required (Roberts & Bradley,
Collaboration Challenges in Community Telecommunication Networks

1924. However, planning follows the visioning that the leader has enacted.

**Challenges in the Formation Phase**

**Defining and Agreeing on the Transformation Effort**

It is argued that a multi-sectoral and multi-organizational network is needed for a transformation to an intelligent community. The wide variety of stakeholders impact the style of leadership and structure needed for joint initiatives. The leader (or leaders in the case of shared roles) shares a vision of a desirable future and initiates a flexible structure that can accommodate differences in orientation (profit versus not for profit), time horizons (short versus long term), and civic engagement (self versus community focus). Given the diversity of stakeholders, the visioning must be consistent and persuasive, but large enough in scope so that stakeholders can personalize the vision to suit their personal and organizational interests. Key activities include:

- Utilizing context issues to create a sense of urgency and sell the concept of the community network
- Identifying solutions to problems and synergistic opportunities
- Preparing a plan for producing meaningful and motivating outcomes

Agreeing on the vision depends on the availability and abilities of the local leader. Individuals with strong communication skills, an established reputation of trustworthiness, an ability to deliver on promises made, and conceptual skills to craft a vision are in short supply. While large communities have a greater pool of candidates, small communities may have to draw more on external talent and work hard on establishing trustworthiness.

**Assembling and Mobilizing Resources**

The community network depends upon its board to acquire the physical, financial, and organizational resources that make a broadband network functional. Collaboration among stakeholders and partners facilitates the pooling of their resources. Choosing board members should flow from resource requirements and the likelihood that the stakeholders recruited or volunteering are favorably disposed to sharing with other organizations. Community citizenship of board members channels the resources to create and enhance the network. Key activities include:

- Assembling the representatives of a variety of public and private sector organizations to form the board, including both small and large stakeholders;
- Mobilizing the resources controlled by board members and reaching out to obtain vital resources from the environment.

Too many resources may harm the development of the network if the board lacks the capability to make good use of them. Waste would damage the network’s reputation and make future resource acquisitions more difficult. Likewise, too few resources can harm the network because the scope of activities would be narrow and appeal to only a small segment of the community’s population. A narrow focus would appear self-serving and lack broad legitimacy.

**Assembling and Mobilizing Trust**

For the board to be effective in creating and enhancing the network’s resource base, its members must trust each other so that extensive sharing becomes possible. When stakeholders engage in joint efforts and initiatives, they are putting the community first and themselves second, making them vulnerable to exploitation by less citizen-
minded organizations. When trust exists on the board, stakeholders can tolerate some exposure. Therefore building and maintaining trust in a realistic manner is essential to the network’s resource base and projects. Key activities include:

• Assembling the board membership on the basis of reputation, prior experience, and diversity of stakeholders;
• Creating a shared vision that reflects the underlying values of community spirit;
• Distinguishing between fragile and resilient trust, and building the latter.

Building and maintaining resilient trust is at the core of the inter-dependent culture that emerges in the network. When a transformational vision is complemented with solid resources and trust, the community network has met the challenges of the formation phase of its development and is ready to shift into the performance phase.

Performance Phase of Development

The performance phase of network development is centred on concrete projects that require the pooling of resources by its members. The resources may be tangible (finances, staff secondment, office space, and equipment) and intangible (time, information, influence, and reputation) in nature. Pooling is facilitated by both the culture and structure of the network in which horizontal interactions, exchanges among equals, are based on trust. These resources are organized and controlled to attain the project objectives, and the management style is collaborative and accountable to the membership of the network. Pursuing these objectives gives collaborators opportunities to learn how they can make the network function effectively. In the short term, the level of attainment of the project’s objectives dominates; small wins and their public recognition are important to confirm the value of the network (Bouwen & Taillieu, 2004). Effective project management is needed. In the long term, the board focuses on the level of attainment of the broad goals of the network. To ensure that the projects and the general management of the network are aligned with the original vision and goals, effective leadership is required. Table 2 provides an overview of the critical factors in the performance phase.

**CONTEXT OF PERFORMANCE**

The interdependence of members within a community network is reflected in both its structure (O’Toole, 1997) and culture. Structure requires

Table 2. Influencing factors at the performance phase

<table>
<thead>
<tr>
<th>Context</th>
<th>Content</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Structure</td>
<td>• Goals</td>
<td>• Team management</td>
</tr>
<tr>
<td>• Roles</td>
<td>• Achievement/outputs</td>
<td>• User appropriation</td>
</tr>
<tr>
<td>• Trust</td>
<td>• Innovation</td>
<td>• Communication</td>
</tr>
<tr>
<td>• Power of stakeholders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Interdependence &amp; Culture</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Collaboration Challenges in Community Telecommunication Networks

careful attention because a poor structure—one that gives too much power to one partner or that does not embody the values of stakeholders—will affect the performance and longevity of the collaboration.

Poyhonen and Smedlund (2004) and Nootueboom (1999) identified three network structures: vertical, horizontal, and diagonal. The latter consists of firms and organizations from several different lines of business. A diagonal structure is appropriate for community networks because it includes as many collaborators as possible to create synergy and innovation within and between sectors; transformational, as opposed to incremental change, is facilitated. The success of collaborative networks is contingent on managing the ambiguity, complexity, and dynamics of the structure. It becomes more important in the performance phase because it must sustain an action plan and organize resources to carry it out. However, a telecommunication network is developed to resolve dynamic context issues and can only do so within a process of continuous improvement. A rigid structure that minimizes innovation diminishes the network’s sustainability. Though difficult to assess, the effectiveness of the structure can be judged by its internal coherence and fit with the culture of the network. This puzzle, identified by Bailey & McNally-Koney (1996), needs a solution that retains the fluidity of communications and decision-making, while providing for a framework for productivity and sustainability.

Collaboration is associated with incremental innovation when partners share on several levels: a larger purpose, explicit and voluntary membership, an interactive process, and temporal property (Roberts & Bradley, 1991). Hardy and Phillips (1998) pointed out that more powerful stakeholders may force collaboration on weaker players to control them. Consequently, there is a lessening of the level of interdependence and common vision. Weaker stakeholders are bound to minimize their participation and find excuses to exit the network when they are being coerced. Though asymmetrical power is a likely reality, leaders that seek innovation must put less emphasis on control and more on incentives and opportunities.

Creating a culture of collaboration gives coherence to the stream of actions that builds the community network. Collaboration is described as a relational system of individuals within groups in which individuals share mutual aspirations and a common conceptual framework (Bailey & McNally-Koney, 1996). Individuals are guided by their sense of fairness and their motives toward others (caring and concern, and commitment to work together over time). Through communication and shared experiences, they create a system of shared assumptions and values, and accepted approaches and solutions to problems, including collective sanctions, to safeguard exchanges and reinforce acceptable behaviors (Jones et al., 1997). Sanctions may include exclusion from certain benefits (present or future) and opportunities (participation in projects), and as a last measure forced exit (temporary or permanent) from the network.

Collaborators often choose to stay in a poorly performing network based on the strength of their social ties. However, if they conclude that they can meet all of their needs outside of the network, they may view the network as superfluous (Brown et al., 1998). Linkages or interdependence must be solid and intentional (Bailey & McNally-Koney, 1996) and may be a strong indicator of sustainability (Olk & Young, 1997). Conversely, Brown et al. (1998) identified that greater resource interdependence makes successful partnerships more difficult to achieve. In order to find common ground and encourage persistence, the reasons for enhancing an interdependence need to be emphasized, and stakeholders must want to belong and believe in the vision.
Content of Performance (Specific Projects)

The content of performance includes a wide variety of projects that meet the goals of the network, including the needs of stakeholders and users. Among them are projects to launch or upgrade an infrastructure, acquire devices to deliver applications, develop content for the network, and promote the network to potential users. The outcomes include cost savings to deliver services, revenues from users, and additional capability for the social and economic development of the community.

Waits (2000) described collaborative networks in terms of their pursuits:

- **Co-inform:** actions to identify members and impacts, promote a heightened awareness of the issues, and improve communication among the members;
- **Co-learn:** educational and training programs sponsored by the network;
- **Co-market:** collective activities that promote member products or services abroad or domestically;
- **Co-purchase:** activities to strengthen buyer-supplier linkages or to jointly buy expensive equipment;
- **Co-produce:** alliances to make a product together or conduct R&D together;
- **Co-build economic foundations:** activities to build stronger educational, financial, and governmental institutions that enable them to compete better.

Some of these pursuits appear easier to realize and only require fragile trust (co-inform and co-learn). They are more likely to give quick “small wins.” Others may be challenging and require resilient trust (co-market, co-purchase, and co-produce); their success will take more time but are more highly valued. Co-building economic foundations appeals less to self-interest and more to a communal interest, and depends on a broad vision that will lead to a series of concrete actions and sustained effort. Waits’ objectives are compatible with each other, but have different time horizons and commitments. The strength of the formation phase influences the commitment of stakeholders in the development phase. In particular, a strong collaborative climate encourages them to be patient and willing to invest additional time and resources to achieve long term goals.

PROCESS OF PERFORMANCE

Leaders require managerial sophistication to recognize appropriate circumstances and tools for collaboration (Glatter, 2004). In networks, collaboration depends upon an ongoing communicative process (Lawrence et al., 1999). Roles and responsibilities are negotiated in a context where no legitimate authority is necessarily recognized (Glatter, 2004; Lawrence et al., 1999; Lowndes & Skelcher, 1998). Like in partnerships, there is concern for trust, politics, emotions, and results. Furthermore, leaders require an understanding of user appropriation of the digital network to effectively channel the collaborative efforts.

Du Gay et al. (1997) describe the appropriation of technology as an active process in which users make choices around the selection and local deployment of the technological components, and create meaning and sense of the technology. Appropriation has both a technical and cultural side. In this spirit, Williams et al. (2005) have argued that user appropriation has two distinct but inter-related processes: innovation (users adjust and innovate to improve the usefulness of the technology) and domestication (users adapt the use of the technology to integrate it meaningfully in their activities). When both processes are fully engaged, the community may be said to have a “digital” culture that sustains the network.
The pace of change within the network must be properly managed. Effective use of communication will allow collaborators to react and contribute. Because of large boards and membership and turnover in representation, some collaborators may not know everyone or their status. Indeed, some may be confused over the degree of autonomy they have in making decisions for their organization (Huxham & Vangen, 2000). Changes in government mandates and organizational priorities create uncertainty as collaborators plan and structure the network. Communication and recognition of accomplishments become important to keep everyone focused.

The board’s effectiveness in tackling problems within their community as well as within their respective organizations will directly influence the achievement of the intelligent community objectives. Leaders need to guide the board and create bridges with important outside players. They must align the requirements of their own organization with the vision of the intelligent community initiative for success; they must create a high performance team environment (Albert, 2005; Wheelan, 1999; Smith, 1994). This standard is not easily achievable, especially for a volunteer board with diverse membership and affiliation.

Challenges in the Performance Phase

Continuing Challenges from the Formation Phase

The consensus on the community vision that was created in the formation phase needs to be reinforced. The leader can remind stakeholders of the urgency to capture opportunities, but must incorporate measures for sustaining collective efforts. Key transformation activities include:

- Expanding planning and monitoring projects and measures of performance;
- Marketing the network concept to mobilize and gain the support of the wider community and further engage the stakeholders.

In terms of resources, the community network continues to depend upon its board to acquire resources to develop, acquire, and develop applications to attract numerous users. Key activities include:

- Modifying board membership to improve the resource base of the network as projects change over time;
- Engaging both small and large partners for innovation to create new resources;
- Creating a small management structure for the performance phase of the network.

As for trust, the performance phase requires continuing sharing of resources in the face of uncertain outcomes. Key activities include:

- Applying different trust standards as the situation warrants;
- Encouraging the broad sharing of resources instead of specialized contributions.

Resilient trust can block new stakeholders and partners from joining the network; they may have key resources but be deemed untrustworthy. In such a case, the network requires the flexibility to resort to fragile trust with its emphasis on formal agreements and contracts. The reverse situation can also damage the network, when fragile trust dominates relationships. While formal contracts increase accountability of the parties, they are narrow in scope and participation is contingent on self-interests being satisfied. Community considerations remain secondary. In time and through active leadership, these new members may buy into community citizenship through association and success.
Balancing Leadership and Collaborative Management

Both the formation and performance phases of development have their champion. The leader dominates the formation (and re-formation) phase through visioning, planning, and attracting and retaining stakeholders with key resources and disposed to collaborate. The manager guides and maintains the performance phase, and ensures that both tangible and intangible benefits are created for the stakeholders and the community.

The “collaborative” manager is needed to reinforce the user appropriation by supporting the innofusion and domestication in which users engage. By encouraging the involvement of intermediaries (e.g., Chamber of Commerce, owner of a cybercafé, entrepreneur who wants to keep control), the network manager allows the network to evolve along lines that reflect the different groups and segments in the community’s population (Williams et al., 2005).

Formal planning becomes less important, as a pattern of coherent projects becomes established. At the same time, these intermediaries (or small groups of individuals in large networks) interact to keep the board informed and reinforce their individual efforts. By working together, they ensure that the vision of the network creates a coherent set of initiatives and projects, and opportunities and issues relevant to the board meetings are identified. Key activities include:

- Encouraging innovation and proper planning to achieve the transformation effort;
- Reaching out to intermediaries to broaden user involvement;
- Ensuring that the vision that binds the board members remains true to the community values as the network develops and expands;
- Confronting head-on the need to modify the board composition to respond to internal or external factors;
- Managing projects with a blend of fragile and resilient trust, the former with binding contracts and the latter with negotiation and concurrence;
- Choosing projects that are likely to succeed and that are valued by the stakeholders;
- Building and maintaining redundant communication systems, both formal and informal, to reflect the culture of inter-dependence that binds the stakeholders of the network.

The network can be damaged by a dominant leader or manager who insists on being involved at all times and on controlling the process, whether at the board or project level. This situation emerges when there is a failure to share multiple roles and to act as a team. The lack of experienced persons may push one individual to assume both the leadership and managerial role; this solution ensures positive momentum, but may block future sharing of roles as the incumbent becomes entrenched. Similarly, the abundance of strong and experienced personalities facilitates the sharing of roles, but may slow down momentum as too many persons insist on prominence. Developing a team spirit among members of the board and management should be encouraged as early as possible in the network’s development (Albert, 2005).

Collaboration Challenges for Users

At the formation stage, the infrastructure and applications are planned and guided by a vision. Stakeholder requirements are addressed in the planning of the network through the methods of user representation. At the performance stage, when the network is functional, the users actualize the network in both expected and emerging ways. A community network is validated by the applications it makes available to its users, and the extent to which the users actually use them. Furthermore, the design features of the telecommunication network influence the collaboration.
opportunities that the network creates. When the network design enhances collaboration, it has succeeded in creating effective socio-technical patterns (Huysman & Wulf, 2005; Evans & Brooks, 2005).

**Challenges for Individual Users**

IT and a community network challenge the individual because they put into question existing ideas and routines, and add knowledge and skill requirements. Being open to change means making efforts to understand and use the network. The younger generation makes more use of the internet than the established older generation for social contact and is likely to push for internet connection in the home (Bernier & Laflamme, 2005; Crowley, 2002). The older adults are more likely to be introduced to ICT changes in the workplace. Age aside, the Internet facilitates the local-global link through which knowledge and expertise from around the world can be channelled to community members (Stevenson, 2002). Creative individuals can interact to exchange expertise and create innovations (e.g., open source development), and are motivated by reputation and recognition built into the Web site (Fischer et al., 2004). To generate ideas, group support systems that ensure anonymity appear more effective (Pissarra & Jesuino, 2005). In general, the individual must learn to assess the trustworthiness of the Internet information sources (Franklin, 1999; May, 2002) and assume risks when making transactions online. Similarly, participating in virtual communities and discussion forums challenges the individual to change roles from spectator to contributor (Ginsberg, 2001) and activist.

**Challenges for Organizational Users**

Organizations that are stakeholders in the community network need to share their “network” vision with their board members, managers, employees, and organizational partners within their supply chains and customer/client networks. Key individuals likely were involved in the network formation stage to ensure that the design of the systems would support expected transactions and activities. At the performance stage, each organization is challenged to mobilize its ICTs, skill base and network use, and do so in dialogue and coordination with their organizational networks. Internally, this means empowering employees and lower levels of management through information systems and decision-making authority. Externally, this refers to the network of relations and the integration of the organizational and community networks. Failure to have extensive collaboration diminishes the benefits that the community network can deliver to stakeholders. Knowledge sharing (Van den Hooff et al., 2004) and knowledge management (Ackerman & Haverton, 2004) are useful frameworks for channelling this collaboration. In addition, involvement can include intra-preneurship (Von Oetinger, 2005) and joint ventures supported by collaborative groupware (McKnight & Bontis, 2002). The organization can also reach out to innovators and entrepreneurs in the community, who view the network as their business platform, and initiate partnerships. The above array of activities pushes leaders and senior managers to adopt an organizational model that incorporates trust.

**Challenges for the Community**

As the community network is fully implemented, the stewardship vision (Block, 1993) incipient in the formation phase must be reinforced by extending inclusiveness to all segments of the local population, imagining a broad culture of use, and providing for economic development with a digital component. Community leaders should have concrete programs to diminish access barriers such as network connectivity at a reasonable cost (or at no cost for public terminals) and to training...
and education. Adoption of the network will vary across socio-economic dimensions, and programs are needed that are adapted to specific groups such as youth, seniors, and the non-profit and small business sectors. Developing and implementing these programs can take place with community stakeholders in collaborative projects. An innovation culture (Martins & Terblanche, 2003), linked to the network, can be encouraged.

A culture of “digital” use is emerging in many communities; the Internet and its many activities are being integrated into everyday routines of social communication, work, and play (Bernier & Laflamme, 2005; Crowley, 2002; Wellman et al., 2001). In contrast, civic participation has had less success. The evidence indicates that internet use reinforces civic participation and makes it more sophisticated, but does not increase the levels of activity (Shah, 2002; Uslaner, 2004; Warkentin & Mingst, 2000; Wellman et al., 2001). Pigg (2001) has argued that networks can be designed to enhance civic participation, but so far, these designs have failed to incorporate the nature of participation. The designs typically focus on customer services and support instead of sharing of information, ideas, and knowledge to influence civic decisions. With a customer focus, the civic authorities may increase the satisfaction of its citizenry, whereas a participation focus obliges the authorities to share decision-making powers and accept more uncertainty in the process and outcomes.

CONCLUSION

A community network faces four inter-related collaboration challenges during its development that are tied to transformation, resources, trust, and management. When these challenges are met, the network will have a solid culture and structure of interdependence, and the flexibility to change over time. The network will maintain a positive momentum that is constructive and manageable, and lead to medium and long-term sustainability. When these challenges are not met adequately, the pace of change will be either too slow or too fast, or blocked at some point in time. Sustainability of the network will be compromised unless the underlying issues are addressed.

These four challenges are anchored in the proposed network development model where formation and performance phases, and adaptation through reformation are critical for the sustainability of the community network. Policy makers and change agents among the stakeholders of community networks are well advised to shape their interventions with the aim of establishing and maintaining positive momentum, while paying continued attention to issues of visioning, resources, trust, leadership, and management. They would do well to expand their views of technology development to include user appropriation and the challenges that users face. They must accept the uncertainty that is inevitable with user involvement to support the goal of network sustainability.

REFERENCES


Collaboration Challenges in Community Telecommunication Networks


Collaboration Challenges in Community Telecommunication Networks


ENDNOTES

1 The authors acknowledge the helpful comments of the reviewers. By addressing their concerns and suggestions, this article found a better balance between organizational and involvement issues.

2 Multi-organizational collaboration, partnerships, and networks are considered interchangeable terms and refer to a variety of organizations collaborating for a common purpose. “Collaborative network” is proposed as an inclusive alternative.

3 A stakeholder is defined as an organization that contributes programs and services to the network. A partner is one that makes a financial contribution to the overall project.

4 The term downloading has become a popular expression in Canada as a result of higher levels of government shifting responsibility for programs to lower levels of government. Municipalities have inherited a number of costs and responsibilities previously held by the province and the province has inherited responsibilities previously held by the federal government.

5 These communities have been highlighted at the annual conference of ICF (Intelligent Communities Forum).

Chapter 5.18
Planning and Managing the Human Factors for the Adoption and Diffusion of Object-Oriented Software Development Processes

Magdy K. Serour
University of Technology, Sydney, Australia

ABSTRACT

Although there are a large number of contemporary software development processes/methodologies available to assist and guide software professionals in developing software systems, there is no specific process that can assist organizations in planning and managing their transition to this new work environment. As a result, there are still a large number of information technology (IT) organizations that have not yet implemented any object-oriented (OO) process. For them, the transition to a new work environment and the adoption and utilization of a software process implies a number of problems, commonly including necessary human and organizational resistance to the ensuing cultural change. This chapter provides IT organizations and professionals with insights into the most important key success factors that may promote the entire process of organizational change. We investigate the effect of various human factors on the adoption and diffusion of an object-oriented software development process. Some of the human factors include motivation, leadership, resistance to culture change, and willingness and readiness to change. In addition, this chapter explores the significant role of these factors in controlling the entire process of implementing an OO process in practice, emphasizing the significance of planning and managing these “soft” factors to achieve clear advantages and gain enviable results.
INTRODUCTION

This chapter investigates and examines the effect of various human behavioral patterns during the organizational transition to an object-oriented (OO) work environment, and the adoption and diffusion of an OO software development process. Technology is only a tool; what makes the difference is the individual who makes use of the technology, and the culture that motivates people to realize and understand the advantages of adopting such technology (Zakaria & Yusof, 2001).

During any paradigm shift, human tendencies play a critical role that may invariably result in either success or failure. Examples of such human aspects may include cultural change coupled with people’s resistance, motivation, education and training, communications, and leadership. Collectively, these factors can form either opposing or supporting forces that may influence and impact on the entire transition process. Therefore, human aspects must be seriously considered, well addressed, planned, and managed for a rewarding result.

Past studies (e.g., Gibson, 1999; Ioannidis & Gopalakrishnan, 1999; Nambisan & Wang, 1999; Auer & Dobler, 2000; Jurison, 2000; Burshy, 2001) of the process of organizational transition have related the transition process to how organizations adopted innovation, ideas, new technologies (e.g., Web services and e-business), or new “ways of doing things” (e.g., the adoption and deployment of an OO process).

What these processes missed in the past was the first (and the most critical) step towards the adoption of a new technology. They all missed the study of moving organizations from their current state or environment to their desired one where they can feel comfortable, familiar, and confident to adopt and diffuse an innovation or new technologies such as OO processes. Getting organizations ready to adopt and diffuse a new technology involves a number of serious managerial decisions that must be made to provide full management support, dedication, and commitment. Organizations must feel comfortable and familiar with the new way of “doing things” before any attempt is made to implement these new ways in practice to avoid or lessen people’s natural resistance to change, and also increase their acceptance and readiness.

Hence, the main objective of investigating the impact of human issues is to gain a full understanding of individual behavior during the transition and also to examine different human factors that influence the response of individuals within organizations toward the adoption of an OO software development process.

ORGANIZATIONAL CHANGE AND HUMAN FACTORS

“The greatest difficulty in the world is not for people to accept new ideas, but to make them forget about old ideas.” John Maynard Keynes

People Behavior During Organizational Change

During an organizational transition, different people play different roles, such as motivators, adopters, resistors, opposers, and neutral or observers (Bridges, 1995). How they respond to change during transition can, and in most cases does, dominate and determine the success or failure of the entire project. The inextricable reality is that people are different, and so act and react to changes differently; from time-to-time even the same person can behave in a different manner.

Bridges (1995) claims that changes are always accompanied by natural resistance, as changes often drive people out of their comfort zone. Consequently, people can develop a resistance to change and become the main obstacle to the whole organizational change.
Once an organization comes to realize what it needs to achieve and decides how it will accomplish its goals, the main challenge becomes the issue of effective and efficient management of human factors. It is quite surprising to know that 80% of project failures are traced back to mismanagement of human factors (Jacobson, Ericsson, & Jacobson, 1995).

Unfortunately, there are many organizations still struggling to deal with difficulties related to the effective management of human or sociological factors during technology adoption and diffusion. This type of problem is usually caused by management’s lack of commitment to the human factors of IT. Szewczak and Khosrow-Pour (1996) relate this problem of mismanagement of human aspects to the fact that, in general, organizations traditionally invest a significant proportion of their resources to obtain the necessary hardware and software technologies, but with insignificant investment in the human aspect of technology. An experienced software development team that has been around for more than a decade, for example, is likely to have superior expertise in (and consequently be comfortable with) traditional software modeling techniques such as Flow Charts, Data Flow, and Entity Relationship Diagrams. Professionals of this type would not be openly receptive to changing their existing work culture and switching to modern OO techniques such as Object Model, Use Case, and interactions Diagrams. This kind of human culture change can form a major challenge during the transition that may increase people’s resistance to change.

The Challenges of Human Factors

In general, human aspects are the most difficult challenge to be addressed during any organizational change (Zakaria & Yusof, 2001). The transition process to OO and the adoption of an OO process usually involves a large number of technical as well as non-technical issues. Various authors have referred to these non-technical issues as soft factors (Constantine, 1995) and sociological factors (DeMarco & Lister, 1987). We call these ‘human’ Key Success Factors in this discussion, since they deal with the ‘human aspect’ of change. These human factors necessarily require complete understanding and managing alongside the technological factors. Since OO contains elements relevant to all stages of the development lifecycle (not only coding), and includes models for requirements engineering, project management, team building, and so on, adopting OO process requires a combination of learning about technical issues as well as larger scale sociological issues.

People, procedures, and tools are the three critical aspects that must be well planned and managed during any organizational change (Serour, Henderson-Sellers, Hughes, Winder, & Chow, 2002). Certainly out of these three aspects, people are the most demanding aspect to be changed and managed. Eason (1983) argues that the human factors of technology are far more important than technical factors. The most challenging aspects of the transitioning to object-orientation remain in establishing a new software development environment and in introducing a new work culture to managers, software developers, and customers (Ushakov, 2000).

As a result, it is imprudent for management to expect every team member to agree with all the proposed changes. It may even be unwise to anticipate full commitment and belief in the new organization mission, especially at the early stage of transitioning. Organizations must face reality by realizing that change is often difficult in the best of circumstances and seldom goes exactly to plan. Management must be mindful that only a minority of individuals will wholeheartedly welcome any proposed changes to the way they have done things for so long.

The challenge to software professionals is to change and adapt to a new environment. A pos-
possible scenario may ensue where developers need to adopt new ways of ‘thinking’ about software, followed by OO modeling/designing, developing, quality assuring, and testing the software. Adopting a new approach of developing software may include technical factors (e.g., CASE tools, Programming languages, and databases) that require a reasonable degree of human culture change in order to utilize them to their maximum potential. Individuals, especially those working on software projects, work under the influence of various personal, motivational, and social factors. These factors, more often than not, remain in the background. Due to their elusive and intangible nature, these human factors are often difficult to discern, analyze, and improve on when change is about to be implemented.

As a result, moving working professionals to new ways of perceiving and undertaking various tasks is very difficult (Fingar, 1996). However, many studies such as Szewczak and Khosrow-Pour (1996) suggest many organizations are still struggling to deal with problems related to the human aspects of technology. Furthermore, they correlated this problem to management’s lack of commitment to the human side of IT.

It is vital, then, that organizations pay careful attention in addressing all human or soft factors, and be open and ready to confront and solve any conflicts that may influence the entire transition. Organizations need to appreciate individual behavior when examining the transition and the adoption of new OO process; they need to investigate and explore the different factors that influence the response of individuals within the organization towards the new technology. To accomplish a successful transition, an organization needs to advance people’s motivation and enthusiasm, maintain management commitment, and provide efficient and persuasive mentors whose leadership may be seen as the source of expert guidance (Jacobson et al., 1995).

The art of managing cultural interfaces has become an everyday business challenge at every organizational level (O’Hara-Devereaux & Johansen, 1994). The main question here, and management’s major challenge, is how an organization makes the transition a driving force for all people involved in being adopters and supporters of the change, instead of opposers or neutral players. Constantine (1996), during the OOPSLA’96 panel discussion pertaining to human factors, contended that it is quite easy to communicate with machines and solve problems, but with people it is difficult. The main reason is because people are very difficult to “generalize.”

**Human Factors and Risk Issues**

Organizations must be aware of the consequences and the possible risks involved in mismanaging the human factors during the introduction of a new work culture such as OO process. The improper planning and managing of human factors can easily lead to undesirable consequences, such as building resistance to change, and adding more confusion, uncertainty, and fear that can considerably diminish the chance of success. Schein (1999) interestingly states that there was no such failure for technology adoption; instead it was a failure to understand the organizational and the individuals’ culture.

As a result, organizations that are aiming to adopt an OO process need to address not only the technological factors of the adoption, but also the human factors. As an example, organizations must provide adequate education and training to their people in order to understand and grasp the fundamentals and underpinning concepts of object-orientation and development processes, as it is a critical factor to avoid those risky consequences. Bridges (1995) argues that the most prevailing cause for the unsuccessful implantation of organizational changes can be attributed
to a lack of planning in managing the impact of change on individuals. He also emphasizes that many organizations that are not properly geared to handle this facet of change can run the major risk of jeopardizing their own existence.

Therefore, management may be very well advised to create a plan to manage and mitigate these potential risk factors by answering the following questions:

- How are people likely to respond to the change?
- What is required to convince people that the change is worth the effort?
- What actions are necessary to earn people’s support and commitment to the change?

In addition, management needs to identify the ultimate and most effective means in achieving the following objectives:

- Selling the change to all people involved, including customers.
- Motivating people to make the transition.
- Reducing people’s resistance to change.
- Eliminating people’s fear and uncertainty.
- Minimizing the change’s disruption to people.
- Mitigating the increasing tensions between people during the change.
- Encouraging people to be enthusiastic for the change, as opposed to being apathetic and an obstacle.

By answering the above questions and achieving the related objectives, organizations will be able to use persuasive approaches to human change, ensuring that everyone is comfortable and willing to accept, and make use of, the new OO process with all its associated changes.

**THE HUMAN FACTORS**

**Human Culture and Necessary Culture Change**

“In a time of rapid change, standing still is the most dangerous course of action.” Brian Tracy

The Oxford English Dictionary (9th edition) broadly defines human culture as the arts and other manifestations of human intellectual achievement regarded collectively as the improvement by mental or physical training. In particular, Palvia, Palvia, and Roche (1996) define the culture of IT professionals as the set of values and practices shared by these members of an organization involved in information technology activities, including managers, developers, and customers/end users.

Personal culture is usually characterized and distinguished by the individual’s values, such as behavior, attitude, experience, and beliefs. There are people who work well under pressure, whereas others work well only when properly supervised and directed, and then there are the ‘cowboys’ (Constantine, 1993) who prefer to work on their own. Cooper (1994) asserts that an IT person’s culture may resist the introduction of a new process which realigns status, power, and working habits, especially when they violate some of the group’s shared values. Current personal culture can be incompatible with certain new processes to the degree that risky consequences may be incurred, including resistance to change, a negative attitude and behavior, implementation failure, or the achievement of totally unsatisfactory results. Human culture change—that is, the physiological change that people undergo to alter the way they carry out their work on a daily basis—is one of the hardest and most longstanding parts of the adoption of an OO software development process (Fayad & Laitinen, 1998).
Natural Resistance to Change

Coupled with the introduction of a new work culture, people tend to naturally build resistance to any challenge of changing their culture and/or learning new things. The unfamiliarity with the new changes can lead to a discomfort that naturally increases people’s resistance. In general, change is often seen as a personal threat by those involved in transition (Huse, 1975).

Resistance to change can also come from management, project leaders, and customers/end users for similar reasons, including fear of change and uncertainty of their capability of carrying out these changes (Fayad & Laitinen, 1998). Furthermore, adopting an OO process also requires people to advance their skills and knowledge, and/or gain new ones, as well as learning new tools and techniques. All these changes can lead to a threat that, if people are not capable of changing their culture, they will be out of the workforce, and be replaced by others who possess the required OO knowledge and skills and are capable of utilizing the new process. Resistance may also happen during the course of adoption when people are faced with serious impediments. For example, people may reach a stage when they feel that they cannot use their old ways (ad hoc) and at the same time they are not comfortable with the new ways (OO process). They then try to escape or oppose the changes. This leads to an increased level of resistance.

During organizational change, managing people’s resistance becomes a critical issue that must be seriously considered so as to accomplish satisfactory results. For that reason, organizations must be able to effectively manage people’s resistance to leading and directing the change process. To do so, management must first understand what resistance really means. What do people really resist? Do they resist the new environment, new technology, or the changes they have to undertake? And finally, why do people really resist? Do they resist for psychological reasons, technological issues, personal concerns, or a combination of all?

What Resistance Really Means?

Naturally, people want to improve and find better ways of doing things. A part of making improvements is causing changes, and changes are always faced with different types of resistance. People’s resistance can be a result of different human reactions that sometimes form obstacles, impediments, hindrances, and difficulties to change. One of the risky issues regarding people’s resistance is that, sometimes, managers see resistance as a sign of laziness, stupidity, or just unwillingness and opposition to change (Fayad & Laitinen, 1998).

In actual fact, resistance can be a sign of people’s disinterest or they could be busy with more pressing issues. Resistance could also be a signal of conflict of interest or contradictory point of views. Moreover, resistance could be a silent request for assistance, more information, or an assertion of different priorities. Resistance can be viewed as an opportunity to gather information and learn better about current and desired state (Bamberger, 2002).

What People Really Resist?

Naturally, people do not like to lose. They do not like their own things to be taken away from them, and that is exactly what people resist. People associate change with the loss of their existing comforts. People do not resist the change itself, so much as they resist the uncertainties, fear, and discomforts associated with it. People, especially those who are confident and comfortable with their existing culture, resist the idea of changing their ways of doing things, and facing the risk of becoming unfamiliar and uncomfortable with the new ways. In addition, every change involves a degree of risk, and people are naturally reluctant to take risks and face the unknown.
Why People Really Resist?

To manage the resistance to change, it is important first to understand the various reasons behind it. Resistance could happen at an early stage of the introduction of the new change and/or at a later stage, during the change process. People resist change when they are unaware of the need to change and, accordingly, are uncertain of the final result. On the other hand, Bamberger (2002) notes that change often meets with resistance because it is seen as a personal threat that leads to fear of failure and rejection. Fayad and Laitinen (1998) relate resistance to the lack of a clear view of the current state and objective goals. They further note that resistance to change often exists as a result of structural conflicts within the organization. Lack of management’s commitment and inconsistent actions with the new change can also elicit resistance.

Bridges (1995) declares that when changes take place, people get angry, sad, frightened, depressed, and confused. These emotional states can be mistaken for bad morale, but they rarely are. Rather they are more likely to be a sign of grieving, the natural sequence of emotions people go through when they lose something that matters to them. People resist the loss of competence that they once had and which was associated with their old familiar tasks. Transition is tiring; during the change, people build resistance when they feel unfamiliar and uncomfortable with the new ways. Resistance happens when people feel that they cannot use their existing old ways, and at the same time they are not comfortable and familiar with the new ways. From a different perspective, people build resistance when they feel that the new ways they have to follow can negatively affect their productivity. For example, an inappropriate new process or technique can discourage people to change, as it can lead to a drop in people’s productivity.

Accordingly, with the introduction of a new OO process, people’s resistance to transition and culture change could be a result of one or more of the following human behavior factors:

- They see the transition with its necessary changes as a threat to their jobs.
- They do not have inadequate knowledge and experience related to the new OO process.
- They are afraid of learning new ways of carrying out their jobs.
- They doubt the benefits of adopting a new OO process.
- They are afraid of failure or that they will not be able to understand the new process.
- It is exhausting to learn new things, and some think they are too old to learn a new process.
- Some prefer to do what they know best, even when they acknowledge there may possibly be a better way (The devil you know!).
- They are overloaded with their current projects with no time to learn new things.
- They are not in favor of the new process with its associated modeling language and/or CASE tools.

Even during the transition process, and when changes take place, people can develop more resistance and be less motivated for different reasons including:

- Anxiety rises and motivation falls.
- They are afraid of failing.
- They are unsure of the new way.
- They are afraid that they may be blamed if something goes wrong.
- They try to avoid learning new things.
- They become self-protective.
- They respond slowly and want to go back to the “old way.”
- They doubt the benefits of the new way.
- They feel panic and confused.

Unsurprisingly, even customers may develop some resistance and be reluctant to accept the
changes as a result of the risk factors involved in introducing a new process that would affect their products. Therefore, customers must be involved in the change process owing to their effective role in supporting the organization’s adoption of a new OO process. The customer role needs to be changed from being just a customer requesting and running a software application to that of being an effective and supportive partner in the whole process of producing their products.

Defeating People’s Resistance to Change

As discussed above, many managers see resistance as a sign of laziness, stupidity, or just plain perversity on the part of employees (Fayad & Laitinen, 1998). Lawrence (1969) suggests that resistance to change should not be treated as a problem, but rather as an expected symptom and an opportunity, or a request, for learning and better knowing the unknown. Also, Senge (1990) stated that resistance to change generally has a real basis that must be understood in order for it to be dealt with. In order to manage people’s resistance to change, organizations need to understand better the very human reactions they face when they are asked to do something differently or change their work habits. Management must plan and practice some strategies to manage and deal with people’s resistance.

To do so, management must begin to understand what “resistance” really is? What are the major reasons for people to build resistance to change? Then they must establish a suitable work environment for people to undergo the required changes. Management cannot change people, they have to change themselves and they only change when they have the appropriate and supportive environment (Boyett & Boyett, 2000). Bridges (1995) reported that most managers and leaders put only 10% of their energy into selling the problem, but 90% into selling the solution to the problem. Management must put more energy into “selling” the problem that is the reason for the change, because people are not interested in looking for a solution to a problem they do not know they have. Management must realize that culture change is very tiring. People will feel tired, overwhelmed, down, and depressed. Management must look at those symptoms as a natural human reaction, and work hard to rebuild its people self-confidence and give them the feeling of competence in mastering the new way. Management must realize their people’s capabilities and give them what they can do best without any fear of failure. Anxiety is natural, and the best way to defeat it is to educate and train people on their new environment. Adequate education and training regarding the newly adopted process can easily eliminate the fear of the unknown, the uncertainties about the final result, and thus lead to elimination of people’s resistance.

Defeating Resistance with Participation

One of the most effective ingredients to defeat people’s resistance to change is by encouraging them at an early stage to participate in planning for the change; Huse (1975) confirms this fact by assuring that people—during transition—see change as a threat unless they have participated in its planning. Lawrence (1969) has demonstrated through one of his case studies, where an identical change was introduced to several factory groups, that the first group, who was not offered any explanation, resisted all management’s efforts, whereas the second group, who was involved in the change planning, carried out its transition with minimal resistance, and an initial small productivity drop was rapidly recovered.

Involvement of people in planning their change allows them to understand why they need to go through it and what they should expect. Humphrey (1995) affirms that people’s resistance can be gradually changed to acceptance, once people are convinced of the necessity of the changes, and also they can see the value of undergoing
the change. So, management must show and convince people that the change they need to go through is not a threat, but rather an opportunity for improvement.

Defeating Resistance with Small Wins

Humphrey (1995) asserts that people’s resistance to change is proportional to its magnitude. Therefore, resistance can be reduced by planning a number of small changes instead of a hefty change. Transitioning an IT organization to OO environment and adopting an OO process involves outsized changes. Psychological, organizational, and technological changes that can be planned in an incremental manner are “small wins.” By introducing the new changes in small increments, each increment will be easy to sell and implement. People will feel confident and positive every time they successfully achieve one increment, and become even more enthusiastic and motivated to implement the next increment. This technique can lead to a smooth transition by enhancing people’s willingness to participate and reducing their total resistance. For example, the process of adopting a new OO method could well be started by addressing the Requirements Engineering (RE) activity as a major focus to engender everyone’s involvement. An initial RE approach, using a well-defined technique such as the use case technique, can be introduced and utilized without the burden of all other activities and techniques. Once people feel confident in carrying out the RE activity, another major activity such as user interface design can be introduced in the same manner and so on for all other activities.

Education and Training (Knowledge and Skills)

Younessi and Marut (2002) define education as an opportunity to learn and ideally master a number of theories—principles that enable the recipient to assess, analyze, and act appropriately to a broad range of relevant situations. The impact of education is usually more abstract and wide in scope. They also defined training as the provision of an opportunity to learn and ideally practice some skills in a controlled environment to carry out a particular task(s). The impact of training is usually more focused, direct, and narrow in scope. In other words, education provides people with the answer to “know what,” whereas training provides them with the answer to “know how.” People can gain knowledge and skills either through education and training courses provided by their organization and/or through real-life work experience.

In general, people need education and training in their discipline to enable and empower them to assess and perform their duties in a professional and satisfactory manner. In the context of the adoption of a new OO process, technical people need to learn about the new process and associated technology to feel comfortable when using it (Zakaria & Yusof, 2001). The individual’s knowledge and experience related to the new process play an effective role in making the decision for the adoption. People with adequate and appropriate knowledge and experience of their new OO process can be more self-motivated and enthusiastic to make a transition than others. On the other hand, during an organizational change, lack of knowledge and/or experience could elicit people’s resistance to change and increase their feeling of incompetency and frustration.

Necessity of Education and Training for Transition

Younessi and Marut (2002) have emphasized the vital role of education and training during the adoption of a new OO process by suggesting that education and training is an obligatory component for the successful introduction of software processes into an organization. Furthermore, they considered education and training as a Critical Success Factor for adopting these processes. Perkins and Rao (1990) stated that
the more knowledge and experience people have, the more they are able to contribute to decisions regarding the adoption of OO processes. They also emphasized the impact of training related to the new technology that increases their ability to adopt, diffuse, and master processes, techniques, and tools. Conner (1992) demonstrates the imperative role of education and training in adoption by saying that people change only when they have the capacity to do so.

People’s experience towards OO processes should have a positive impact on the adoption (process) of such technology. In the context of adopting a new OO process, Sultan and Chan (2000) stated that the greater the knowledge and skills of individuals, the more likely they are to adopt it.

Management must recognize the critical nature of proper education and training, since experience has shown that between 25 and 40% of the total cost of an extensive project will be spent on education and training (Mize, 1987). More experience enables people to contribute more towards the transition and the adoption of new OO processes. People’s appropriate knowledge, experience, and education may have a positive impact on their transition. The greater the work experience of individuals with the organization, the more likely they are to transition and adopt the new process. Highly skilled staff can manage themselves more easily, particularly during a paradigm shift. Therefore, an individual’s satisfactory level of knowledge and education towards the new technology and processes forms another imperative management challenge.

Motivation has a remarkable power and is a major influence on people’s behavior towards any achievement (Humphrey, 1997). In order to investigate the importance of motivation during the organizational adoption of OO processes, we need to understand what motivation is, its role during the adoption, and what motivates people.

What is Motivation?

Bolton (2002) defines motivation as a sociological concept used to describe individual factors that produce and maintain certain sorts of human behavior towards a goal. Hence, motivation, as a goal-directed behavior, is a driving force that stimulates people to achieve a set of planned goals.

Motivation is often driven from a desire or an inspiration to accomplish a defined objective, combined with the ability to work towards that objective. In other words, Motivation is the ability of taking good ideas or serious changes, and coupling them with appropriate knowledge and skills to achieve desired objectives.

Consequently, people who are aiming to achieve a specific goal must be both motivated and have the power and ability to work towards that goal. People who are motivated towards an accomplishment must be also capable and empowered to do so (carry out the work).

In the context of this chapter, motivating people during the adoption of a new OO software process could mean those factors which cause individuals within the organization to do more than they otherwise would. For example, to transit a development team to a totally new OO environment and/or adopting a new process, people need to work more than usual to change their existing work culture and adopt a new way. Then, motivation becomes a measure of an individual’s level of readiness and willingness to participate effectively towards a successful transition.

"Motivation is the release of power within a person to accomplish some desired results. It is a major key for achievement and success.” Dr. Len Restall
Role of Motivation During Adoption

Pfeffer (1982) states that when people understand and accept a goal, and believe they can meet it, they will generally work very hard to do so. Additionally and from a different perspective, Maslow (1970), when he described his “Hierarchy of Needs,” stated that people are capable of achieving their goals if they believe that they can achieve them.

Motivation provides people with the understanding of the reasons for the change that increases their acceptability, which in turn improves their ability to accomplish a successful mission. Moreover, motivated people more often than not are capable of defining and achieving their own goals.

During an organizational change, people need to have compelling and persuasive reason(s) why they have to go through changes. Once they are convinced and believe in their mission, they will feel more competent to carry out their necessary changes successfully. This becomes a positive motivation that makes people desire to accomplish their goals that result in their very best performance (Humphrey, 1997).

On the other hand, lack of motivation during adoption can lead to negative consequences that may contribute to undesirable results. Those consequences may include fear, confusion, frustration, and uncertainty. Humphrey (1997) confirms the vital role of motivation by saying, “Without motivated and capable employees, no technical organization can prosper.”

As a result, management must improve and maintain people’s motivation to help them to be more effective and efficient in moving to their desired work environment with the adoption and utilization of a formal OO process. Motivation addresses the degree to which people want to, and are willing to, complete the work necessary to change their existing work culture. Furthermore, during transitioning to a new work environment, management must maintain people’s motivation if they try to give up easily, or too soon, so as to encourage them to keep trying as long as they believe in their goals.

What Motivates People?

Bolton (2002) suggests that a good first step towards understanding what motivates people is to know what people want from their jobs. The answer could be gaining financial advantages, acquiring more skills and knowledge, or working with the latest technologies. In reality, it is very difficult to predict and judge on people’s desire because it depends on the individual’s values and beliefs.

What motivates people to change their work culture and adopt a new process differs, all depending on their needs and perception of the new process. Hence, not all people can be motivated by the same things and to the same degree. Therefore, management—especially during transition—needs to understand why their people do their work, and consequently elicit what motivates them and how to maintain that motivation through the entire process. Understanding people’s aspiration can help management not only motivate people, but also support their motivation by maintaining and increasing the reasons for motivation.

Motivation is often seen as the driving force that moves people to perform some actions. Then, in the context of the adoption and diffusion of an OO process, people need a comfortable, familiar, valuable, and convincing driving force or “motivation factor” to move them to perform more effectively than they usually do.

People feel comfortable when they confidently know how to do their jobs. Enhancing people’s skills and knowledge to the required level for transition makes them feel comfortable with the new process. Formal and professional education and training pertaining to OO processes are considered to be effective and efficient techniques.
Human Factors for the Adoption and Diffusion of OOS Development Processes

In order to motivate people, training must include clarification of language and jargon or commands used to avoid further frustration and anxiety, as new technologies usually have a mystifying and alienating potential. The more comfortable people feel with the new technology (here OO process), the more willing they are to experiment with it (Zakaria & Yusof, 2001).

Motivation factors must also add extra values for people such as financial reward, or learning a new language, process, or tool. For example, someone likely to retire in a year is unlikely to learn a new way of doing things. On the other hand, an enthusiastic newcomer to the project is likely to put in the extra effort needed to pick up new methods of thinking and modeling.

People who are in the move (e.g., changing profession, changing company, or reaching retirement) will not be interested in changing their culture, and they become difficult to motivate to change their current culture because they will not gain any benefits. To motivate people to change their culture, they need assurance of benefiting from the change.

The driving force must also be convincing, it must be relevant to what people usually do, and it must be in a way that people can believe and make use of. Most people must be convinced of the need to change before they will willingly comply (Humphrey, 1995).

Leadership

“Leadership is the ability to provide those functions required for successful group action.”
Weldon Moffitt

Leadership is well thought-out by many researchers and authors as a fundamental activity of project management, and they simply describe it as motivation plus organization (Stogdill, 1974; Thite, 2001; Phillips, 2002; Castle, Luong, & Harris, 2002).

In general, leadership plays a significant role in how people effectively perform their jobs. Rogers (1995) defines leadership as the degree to which an individual is able to guide, direct, and influence other individuals’ attitudes informally in a desired way. Stogdill (1974) describes leadership as a process of influencing group activities toward goal setting and goal achievement.

Leadership is a practice that needs special skills and talent to motivate people to get things done in a favorable and constructive way. Bridges (1995) affirms that leading a team of professionals efficiently is an invaluable resource to any leader who is eager and willing to understand how to inspire team members. Also, it is a vital ability to augment teams’ cooperation and individuals’ commitment and productive participation. Above all, an effective leadership is a Critical Success Factor on how team members become adaptable to changing circumstances (Thite, 2000).

Leadership is basically situational. Fiedler (1967) suggests that there is no one style of leadership that suits every project situation. Rather, it should be contingent upon the nature of the organization, the team, and the project situation. Leadership style should be flexible and agile enough to be adapted to suit the project at hand in the most appropriate manner. Phillips (2002) claims that the “appropriate” leadership style is the style with the highest probability of success. Additionally, a good leader should effectively mix and cooperate with team members.

Leadership Style for Adoption

“A leader is best when people barely know that he/she exists.” Whitter Bynner

During an organizational change situation, leadership style must be adapted to promote and support the organizational change. Social factors such as the style of leadership can influence the individual’s ability to transit to a new work climate. Thite (2000) emphasizes the importance of
the appropriate leadership style during a change situation by considering it as a Critical Success Factor.

Changing people’s work culture and introducing them to a new way of doing their job is usually accompanied by increasing anxiety, ambiguity, and insecurity. During such time, a supportive leadership style can effectively contend with and overcome these problems.

Due to the fact that an effective leadership style should depend on the follower and the project situation, leaders must use the most appropriate style of leadership that best suits the project situation that yields the best chance of success (Phillips, 2002). Furthermore, Phillips states that managers must be able not only to determine the most appropriate leadership style, but also to apply that style correctly.

Stodgill (1974) supports that argument by saying, “The most effective leaders exhibit a degree of versatility and flexibility that enables them to adapt their behavior to the changing and contradictory demands made on them.” Moreover, Hersey, Blanchard, and Johnson (1996) proclaim that successful and effective leaders are able to adapt their leadership style to best fit the requirements of the situation. Therefore, leaders within an organization about to introduce a major change to their work environment—such as the adoption of a new OO process—must have adequate resources (mainly time), a strong interest in the mission, and a greater exposure to the new process they are about to adopt. Sultan and Chan (2000) firmly assert that the greater the opinion leadership among group members, the greater the chance of a successful transition and adoption.

Leadership always means responsibility. Consequently, an effective leadership style for a transitioning organization must be driven from top management with full support and commitment, and it must be also coupled with a rational degree of authority. Co, Patuwo, and Hu (1998) support that by saying: “Ideally, the team leadership should come from top management: one who is a ‘doer’, and who commands respect throughout the entire company.”

During the organizational transition to a new work environment, leadership should foster strong alignment with the organization’s vision. Team members that depend primarily on strong alignment with their organization’s vision are at their best with well-understood practices applied to well-understood problems. Furthermore, successful leaders must be able to create a new work environment in which team members feel comfortable and familiar, and also enjoy what they are doing.

Lockwood (1991) and Constantine (1994) affirm that the working culture shapes the style in which software development is carried out. Different styles of organization during different situations require somewhat different forms of leadership style and management, and each will tend to have somewhat different software practices.

Leaders who are leading their teams through a serious change—such as adopting a new OO process—must be able to:

- Define a clear vision and mission.
- Provide convincing and compelling reasons for the change.
- Focus on and be concerned with the why to aspect of the transition, rather than the how to.
- Manage and motivate team members.
- Maintain an authority level to make appropriate decisions in the face of uncertainty.
- Provide full support and commitment to their followers.
- Provide adequate and appropriate resources including time, education, and training.
- Establish clear communication channels with team members.
- Monitor and review tasks allocated to team members.
- Recognize and appreciate achievements, and reward people for doing good work.
Perception of OO Processes

While IT organizations perceive new OO processes as an essential means to advance their existing software development culture and effectively compete in the marketplace, individuals could have a different perception. Since people are different, their needs and expectations are also different. The more positively people perceive their new process with regard to its characteristics—advantages, observability, compatibility, complexity, and trialability—the more likely it is that the process will be adopted and utilized.

Management of Expectations

From past projects, it has been proven that people with life experience are able to deal with conflicts of interest that may arise due to their ability to make any decision regarding the transition (Hill, Smith, & Mann, 1987). To emphasize the vital role of education on conflict resolution, Barclay (1991) stated that any organization’s members who have adequate education and significant real-life experience are well prepared to deal with interdepartmental conflict. It has to be made very clear from the beginning, to both managers and developers, that following a new OO process is not a magic wand or a silver bullet to make the entire organization’s dreams come true, but rather is simply today’s best approach option for software development. Nonetheless, it is also important that managers appreciate not only the benefits of adopting a new process, but also become aware of all the pitfalls and consequences of changing people’s existing work culture.

Different stakeholders such as managers, developers, and customers may look at, assess, and evaluate OO processes, as a new technology to be adopted and utilized, from a different perspective. Senior managers are always looking for the dollar value return for their spending and investments. They evaluate the technology based on how much it will reduce the cost of software production with improved quality. Project managers and team leaders may assess the new OO processes from a different perspective. They emphasize project management aspects, gaining more control and meeting customer’s expectations with on-time delivery based on building the software system from the pre-tested and proven software components. Software developers may value the proposed process for adoption as the new fad and trend to developing software. They view the transition process as a good opportunity to acquire new skills, and learn new tools and techniques. Unquestionably, this will insert desirable additions in their resume, add value to their worth in the market, and make them in more demand.

From the above discussion, unrealistic expectation from adopting a new OO process and also any conflict of people’s interest can have a negative impact on the entire transition process. People will be reluctant to work effectively together as a team if they do not share the same interest and/or have different expectations.

In order to overcome the barrier of the conflict of interest, management and champion teams must include a specific plan to resolve that conflict between people, and reach mutual perceptions, consensus, and understanding.

Proper education, mentoring, and open discussions are examples of good techniques that can be used to achieve good understanding and realistic expectations of the new process. Stakeholders must come to some sort of consensus as to what they can expect from adopting a new process and also should be fully aware of the new technology pitfalls as much as its benefits. Stakeholders’ consensus and shared understanding, as well as realistic expectations of the new process, can lead to a good working environment that can positively impact on the process of transition and thus the quality of the software produced.
Sharing the Vision with the Organization

People perform better if they feel that they share the value of the change with their firm. Managers need to provide the right organizational climate to ensure that their employees can see that, by working towards the organizational goals, they are also achieving some of their own goals. These goals could be financial rewards, or personal rewards such as the respect of their colleagues, or job satisfaction, or a combination of any number of things that the employee considers to be important.

During organizational change, such as the adoption and diffusion of an OO process, to gain people’s acceptance, management should get people involved during the initial stage, listen to what they have to say, respect their view, and involve them in discussions and debates regarding the change. Follow up with them on any further progress. Give them the feeling that they are the owners and supporters of the change. Give them the impression that we are only interested in the change if it will help them and make their job more efficient and more enjoyable. Never blame the people for any faults or mistakes, but rather blame the product, the techniques, and/or the tools being used. Organizations need to try everything they can during the adoption of a new process to gain their people’s willingness and support such as:

- Choose a flexible software development process/methodology to be tailored to best suit the organization’s environment.
- Obtain senior and middle management support and commitments.
- Set an appropriate change management plan.
- Establish a transition support team to sell the changes to everyone and to deal with all transition issues.
- Plan for adequate resources.
- Plan for rewarding people behind the transition.
- Carry out the changes gradually and avoid overloading people.
- Plan for the appropriate type of education and training to enhance people’s skills that minimize their fear and confusion.
- Introduce new ways in formal and informal ways as required.
- Never blame people for any failure; instead blame the process (no criticism).
- Listen to people and encourage their individual contribution.
- Reinforce the new way of following the new adopted process.
- Start with small jobs with high chance of success so they to gain self-confidence.
- Celebrate success with people.

Organizations should set a well-defined set of guiding beliefs that the mission of the transition and adopting a new OO process is worthy and must be shared with people. Individuals must be in harmony and agreement with an organization’s value and goals. When individuals share the same values and beliefs as the organization, they form a psychological union as a group. They become more congruent to the organization. Thus, the efforts of adopting a new process are determined and rigorous. The sharing of ideas and commitments through cooperation and harmony within an organization often leads to an effective change and positive participation. Needless to say, whenever individual goals are in alignment, or close to alignment, with the organizational goals, then any initiative undertaken by the organization to adopt and diffuse a new technology—in our case, OO process—will be easily accepted and utilized by individuals. When people share the vision and beliefs with their organization, they feel that they share the ownerships of their new process and the accountability of the whole transition procedure.
This feeling can motivate the individuals and assist management to defeat resistance to change should it arise. A strong organizational culture is usually measured by the way in which the key values are intensely held and widely shared. It has a greater influence upon employees than weak culture.

**Communication Channels**

When Zakaria and Yusof (2001) analyzed the implementation of technological change from a user-centered approach, they asserted that communication is a critical element in ensuring the success of a new technological change such as the adoption of OO processes.

Communications are defined as a process by which individuals exchange information through a common system of behavior. The process of communication between individuals has an influence on organizational transition, because it has a strong impact in conjunction with the leadership. A strong and effective communication network has a strong positive impact on adoption of a new technology. Proper and open communication channels can, from one side, help management to express their commitment and support to their people. From the other side, these channels can give people opportunities to discuss and exchange their concerns regarding their transition in order to adopt the new process. This can strongly increase people’s cooperation, which enhances their ability to carry out the required changes effectively. From a project management point of view, open communications keep management in continuous contact with their people, and they can closely observe the transition progress to quickly resolve any problems that may occur. Communication can take different forms such as meetings. During an organizational transition to a new work environment and adopting a new process, meetings—driven by management—are seen as ways of undoubting management’s commitment and support to change.

**CONCLUSION**

It can be clearly noticed that by considering and managing the human factors in great detail, the chances of successful adoption and diffusion of a software development process are enhanced significantly. Based on our research projects and several empirical studies, we conclude the following:

- Human factors play a vital role during the organizational transition to adopt and diffuse an OO process, as they can form either a promoting or resisting force.
- Understanding of the new desired OO process work environment can assist management in putting together a plan to manage the necessary cultural change.
- During an organizational change, managing people’s resistance becomes a critical issue that must be seriously considered so as to accomplish satisfactory results.
- Organizations must be able to effectively manage people’s resistance to leading and directing the change process.
- Involving individuals in planning and making decisions can positively eliminate their resistance to change and enhance their ability to carry out a successful transition.
- The individual’s knowledge and experience related to the new process play an effective role in making the decision for the transition.
- More knowledge and experience enable people to contribute more towards the transition and the adoption of new technologies, including OO processes.
- The greater the work experience of individuals with the organization, the more likely they are to transition and adopt the new process.
- Gaining adequate knowledge and proper training on OO processes and associated
tools can significantly contribute to enhancing people’s ability to positively involve themselves in the transition process.

- The more knowledge and experience the organization’s individuals have regarding OO processes, the more likely they are to change and adopt the new work culture and thus enhance the chances of a successful transition to a totally new OO software development environment.
- When people understand and accept a goal and believe they can meet it, they will generally work very hard to do so.
- Motivation provides people with the understanding of the reasons for the change that increases their acceptability, which in turn improves their ability to accomplish a successful transition.
- Maintaining and enhancing people’s motivation during transition can be a driving force for people to wholeheartedly and positively participate.
- Successful and effective leaders are able to adapt their leadership style to best fit the requirements of the situation.
- The greater the opinion leadership among group members, the greater the chance of a successful transition and adoption.
- The more positively people perceive the new process with regard to its characteristics—advantages, observability, compatibility, complexity, and trialability—the more likely that the process will be adopted and effectively used.
- Management of expectations of the new process and conflict resolution can lead to a good working environment that can positively impact the change process.
- When individuals share the same values and beliefs with the organization, they are more likely to effectively participate towards the change.
- When people share the vision and beliefs with their organization, they feel that they share the ownerships and the accountability of the whole change practice and thus their new OO process.
- The sharing of ideas and commitments through cooperation and harmony within an organization often leads to an effective change and positive participation.
- The more individuals find values in the organizational change, the more likely they are to contribute to the transition success.
- The more congruent and rewarding the individuals perceive the organization’s values, the more positive is the influence on the transition and the adoption.

REFERENCES


*This work was previously published in Management of the Object-Oriented Development Process, edited by B. Roussev and L. Liu, pp. 247-277, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 5.19
Product Customization on the Web:
An Empirical Study of Factors Impacting Choiceboard User Satisfaction

Pratyush Bharati
University of Massachusetts, Boston, USA

Abhijit Chaudhury
Bryant University, USA

ABSTRACT
Choiceboards are Web-based systems that allow consumers to customize their orders. The study investigated factors that affect consumers’ intention to use choiceboards. The research is based on Masons’ theory and DeLone and McLean’s model of information system use. It was found that intention to use is affected by overall satisfaction. In turn, these two factors are positively impacted by factors such as system quality and information quality. In spite of support from theory, the evidence for the factor, information presentation was weak.

INTRODUCTION
E-commerce is coming of age (Markillie, 2004). Sales in the year 2003 exceeded $55 billion, and revenues in 2004 are expected to be at least 20% higher (Syre, 2004). The total impact of e-commerce, however, cannot be expressed in simple sales figures; rather, it lies in changing consumer behavior. Increasingly, consumers visit the Web site of a company to familiarize themselves with the firm’s offerings and prices before deciding to buy. A Web site is becoming the gateway to a firm’s brand, even in the case of off-line firms. Companies that realize the importance of their Web sites use technologies such as e-mail, FAQ, online customer support, bulletin boards, and search
Product Customization on the Web

engines to assist customers in the buying decision process and, obviously, to persuade a purchase of their product.

The choiceboard is a recent addition to this repertoire of technologies, aiding consumers in the decision-making process (Andal-Ancion, Cartwright, & Yip, 2003; Bharati & Chaudhury, 2004a; Collins & Butler, 2003; Liecht, Ramaswamy, & Cohen, 2001; Slywotzky, 2000). A choiceboard is a system that allows customers to design their own products by choosing from a menu of attributes, components, prices, and delivery options (Slywotzky, 2000). For example, in the automobile industry (buyatoyota.com), users can “build” or customize a Toyota and then follow up with a local dealer. In the construction industry (kitchens.com), users can get help to design a kitchen and actually place an order. In the apparel industry (acustomtshirt4u.com), users can select color, fabric, and a suitable logo and lettering. In the entertainment industry (www.apple.com/itunes), customers at the itunes music store can build customized CDs by selecting individual tracks from existing CDs. Finally, in information technology, the Web sites of most computer firms (e.g., www.ibm.com) present individuals with a basic configuration defined by a processor and then flesh out the full configuration with choiceboards, offering hard-drive size, memory, and add-ons such as CD/DVD drive, monitors, and printers.

Although choiceboard technology is being used widely to enhance the customer’s experience, very little is known about the actual impact of this technology on overall user satisfaction or on the intention to use the choiceboard. Similar concerns have been expressed for Web-based decision support systems (Bharati & Chaudhury, 2004b). In particular, it remains unclear how the provision of more information, facilitation of decision making through what-if analysis, and choice comparisons through the use of choiceboard technology affects user satisfaction and the intention to use. In this research, the relationships are developed and operationalized between system-level factors (i.e., quality of the system and information in choiceboards and presentation of information) and user’s decision making and interface satisfaction. Furthermore, the analysis investigates the relationship between information and decision-making satisfaction, with overall satisfaction and intention to use. The statistical analysis consists of path analysis, assessing a pattern of predictive relationships among the measured variables. This research employs the Structural Equation Modeling (SEM) technique to analyze the data and then to assess the pattern of predictive relationships.

The research views information systems’ success in the new domain of e-commerce and, in particular, in the context of choiceboard systems. It attempts to understand how choiceboards facilitate user decision making in the Web-based environment. It then develops a conceptual model that relates system-level factors, user satisfaction factors, and use factors. Specifically, it investigates interrelationships between components of user satisfaction (i.e., interface satisfaction, decision satisfaction, and overall satisfaction) and their combined impact on intention to use.

LITERATURE REVIEW

The research is related to multiple theories such as the consumer decision-making model (Mowen, 1995), consumer information processing model (Bettman, 1979), cognitive decision-making model (Simon, 1955), and information systems (IS) success model (Delone & McLean 1992, 2002). According to Mowen (1995), a consumer transits through several phases, such as problem recognition, a search for alternatives, and an evaluation of alternatives, before making a choice; that is, there is an information-processing phase and then a decision-making phase. In this process, a consumer tries to minimize the cognitive effort required to make a decision and yet maximize
the quality of the decision reached (Bettman, 1990). Furthermore, Bettman (1990) suggests that because of bounded rationality constraint (Simon, 1955), consumers actually will trade off decision quality for a reduction in information processing effort.

Consumers employ decision aids such as calculators, spreadsheets, consumer guides, and Web-based comparison pricing in order to lessen the impact of bounded rationality constraints on decision quality. E-commerce retailers are incorporating choiceboards on their Web sites in order to assist customers in several phases of the decision-making process (Bharati & Chaudhury, 2004a, 2004b). The information search phase, for example, is facilitated by easy revelation of product alternatives, and the decision-making phase of alternatives evaluation is made easier by price and feature comparison. The IS success model (Delone & McLean 1992), with its focus on issues relating to information processing and decision making and its previous research on Web-based DSS (Bharati & Chaudhury, 2004b), is useful in investigating the role of choiceboards in assisting users. In the recent literature, this model has served as the basis for investigating similar research areas such as IS and service quality (Bharati & Berg, 2003). The research on quality of information systems services (Jiang et al., 2000, 2002; Kettinger & Lee, 1997, 1999; Pitt et al., 1995, 1997; Van Dyke et al., 1997, 1999; Watson et al., 1998) and WebQual (Loiacono et al., 2002) also has attempted to investigate this topic in a slightly different way.

Communications theory (Shannon & Weaver, 1949) was illustrated and modified in Mason’s (1978) work to show that classes of information output are at the technical level, semantic level, and influence level. The IS success model (Delone & McLean, 1992, 2002) expanded the concept of levels of output in order to illustrate stages within those levels. Information is communicated to a recipient who either is influenced or not; he or she then impacts organizational performance.

In other words, the information flows from its production to influence the individual and then the organization.

System quality and information quality both singularly and jointly impact use and user satisfaction. This research model is based on the IS success model and employs some of the constructs of that model, specifically at the technical level of system quality and information quality, in the context of choiceboards, and in their impact on different components of user satisfaction (interface satisfaction, decision-making satisfaction, and resultant overall satisfaction). User satisfaction then influences the intention to use. The next section explains the research model and hypotheses.

**RESEARCH MODEL AND HYPOTHESES**

The research model (Figure 1) shows that system and information quality and information presentation impact the different components of user satisfaction and then intention to use. The various constructs and the resulting hypotheses of the model are explained in this section.

**System Quality**

System quality is the individual perception of a system’s overall performance, which is itself a manifestation of system hardware and software. Ease of use (Belardo, Karwan, & Wallace, 1982), convenience of access (Bailey & Pearson, 1983), and system reliability and flexibility (Srinivasan, 1985) are measures employed for the service quality construct.

**Information Quality**

The user estimates the value of an information system after evaluating the quality of information it provides (Gallagher, 1974). Information accuracy (Bailey & Pearson, 1983; Mahmood,
Figure 1. Model with results

1987; Miller & Doyle, 1987; Srinivasan, 1985), completeness (Bailey & Pearson, 1983; Miller & Doyle, 1987), relevance (Bailey & Pearson, 1983; King & Epstein, 1983; Miller & Doyle, 1987; Srinivasan, 1985), content needs (Doll & Torkzadeh, 1988), and timeliness (Bailey & Pearson, 1983; King & Epstein, 1983; Mahmood, 1987; Miller & Doyle, 1987; Srinivasan, 1985) are the measures employed in the information quality construct.

**Information Presentation**

In information presentation, the display of information based on formats, colors, and graphs vs. tables is examined (Vessey, 1994). The interface evaluation has included presentation, format, and processing efficiency characteristics of the interface (Swanson, 1985-1986). The measures used for information presentation construct are graphics, color, presentation style, and navigational efficiency (Swanson, 1985-1986).
Interface Satisfaction

The quality of the information system interface is measured in interface satisfaction. The indicators used to measure interface satisfaction construct are easy to work (Doll & Torkzadeh, 1988; Goodhue, 1990), useful format (Doll & Torkzadeh, 1988; Goodhue, 1990), user friendly (Doll & Torkzadeh, 1988; Goodhue, 1990), does what I want it to do (Davis, 1989; Goodhue, 1990), and clear and understandable (Davis, 1989; Goodhue, 1990).

Hypothesis 1: System quality will contribute positively to interface satisfaction.

Hypothesis 3: Information quality will contribute positively to interface satisfaction.

Hypothesis 5: Good information presentation will contribute positively to interface satisfaction.

Decision-Making Satisfaction

Decision-making satisfaction is the system’s ability to support the user’s decision-making and problem-solving activities. The system’s support to the individual in recognizing problems, structuring problems, and making decisions related to the goal of controlling a business process are part of the construct (Garrity & Sanders, 1998). The construct measures the decision-making satisfaction using decision effectiveness (Chervany, Dickson, & Kozar, 1972) and decision confidence (Goslar, Green, & Hughes, 1986; Guental, Surprenant, & Bubeck, 1984; Zmud, Blocher, & Moffie, 1983).

Hypothesis 2: System quality will contribute positively to decision-making satisfaction.

Hypothesis 4: Information quality will contribute positively to decision-making satisfaction.

Hypothesis 6: Good information presentation will contribute positively to decision-making satisfaction.

Overall Satisfaction

Satisfaction is an important and widely used construct in the IS literature. Numerous researchers have modified the Bailey and Pearson (1983) user satisfaction instrument. The construct of overall satisfaction, a result of interface and decision-making satisfaction, was measured using extremely useful system (Sanders, 1984) and satisfactory in meeting user needs (Alavi & Henderson, 1981; Sanders & Courtney, 1985).

Hypothesis 7: Interface satisfaction will contribute positively to overall satisfaction.

Hypothesis 8: Decision-making satisfaction will contribute positively to overall satisfaction.

Intention to Use

Intention to use a system often has been employed as an important measure of IS success (Chang & Cheung, 2001; DeLone & McLean, 1992; Lucas, 1978; Van der Heijden, 2004; Welke & Konsynski, 1980). Possible to use and intend to use (DeSanctis, 1982) have been employed to measure the intention of the user to use the system construct.

Hypothesis 9: Overall satisfaction will contribute positively to intention to use.

RESEARCH METHODOLOGY

The instrument was constructed based on prior research, and most indicator items were adapted or borrowed from previously validated instruments. The survey was first pretested with a smaller sample and then subsequently refined.
Product Customization on the Web

The survey was administered to subjects who were undergraduate and graduate students at two universities. They were selected as subjects, because they were users of or familiar with choiceboard systems. The experiment was conducted in a laboratory setting with PCs running on the Windows operating system. In conducting the experiment, the researchers adopted the following procedure.

First, the experimental procedure was explained to the subjects. Then, each subject was randomly assigned a Web site that employed a choiceboard that allowed the user to configure a product. The choiceboard sites were of a very similar nature, despite being owned by different firms. After configuring a product on the Web site, each subject completed a survey questionnaire. The total sample for the experiment was 192 subjects.

Structural equation modeling (SEM) was used to analyze the data. SEM subscribes to a causal indicator model with the operational indicators reflective of the unobserved theoretical construct. It allows the specification of measurement errors within a broader context of assessing measurement properties. Confirmatory factor analysis, content validity, unidimensionality analysis, reliability analysis, convergent validity, and criterion-related validity tests were conducted to evaluate the model and constructs (Anderson & Gerbing, 1988; Bollen, 1989; Chin, 1998).

DATA ANALYSIS

Confirmatory Factor Analysis

The measurement properties of the survey instrument were assessed with confirmatory factor analysis. A measurement model comprised of a weighted linear combination of the items in the scale was analyzed. In confirmatory factor analysis, each theoretical construct is specified and analyzed to assess the fit of the data with the measurement model (Ahire, Golhar, & Waller, 1996; Ravichandran & Rai, 1999; Venkatraman, 1989). For constructs with four or more indicators, these guidelines were followed. As some constructs have fewer than three indicators, these constructs were pooled with constructs having four or more indicators. This was done to ensure adequate degrees of freedom for estimation of the model.

Content Validity

Content validity is ensured when the constructs are defined using the literature. The construct should adequately represent and measure the domain of meaning that it is supposed to represent (Bohrnstedt, 1983). If all the items grouped together for each construct reflect the underlying meaning, then content validity exists (Dunn, Seaker, & Waller, 1994). Since there is no rigorous way to assess content validity, in order to ensure thoroughness, multiple items were used to measure the construct (Bohrnstedt, 1983; Churchill, 1979). The instrument employed in the research used several indicators for each construct that were derived from an in-depth literature review, and thus, content validity was ensured (Bohrnstedt, 1983).

Unidimensionality Analysis

A multidimensional construct helps with content validity and is acceptable as long as the scales are unidimensional. A scale has to be unidimensional in order to have both reliability and construct validity (Gerbing & Anderson, 1988). The condition for a unidimensional scale is that the items of a scale estimate one factor. The goodness of fit index (GFI) measures a good fit of the measurement model, as it indicates that all items load significantly on one underlying latent variable. There is no evidence of lack of unidimensionality when GFI is 0.90 or higher for the model. The GFI indices for all the scales are
summarized in Table 1, and the results suggest that all the scales are unidimensional.

**Reliability**

Reliability of a scale is ensured, if the scale is dependable, consistent, or stable (Gatewood & Field, 1990). Cronbach’s alpha coefficient was used to measure reliability, as the items of a scale explain the majority of the variation in the construct vis-à-vis measurement error (Cronbach, 1951). The results indicate that the scale is reliable, because the alpha coefficient is greater than 0.70 (Table 1).

**Convergent Validity**

Considering each item in the scale as a different approach to measure the construct usually assesses convergent validity. This was measured using the Bentler-Bonett coefficient ($\Delta$)(Bentler & Bonett, 1980). The Bentler-Bonett coefficient ($\Delta$) value of 0.9 or above means high convergent validity. All the scales had a Bentler-Bonett coefficient ($\Delta$) of greater than 0.9 (Table 1).

**Criterion-Related Validity**

Criterion-related validity tests the degree to which the outcome is predicted by the constructs.
Product Customization on the Web

Using SEM, the constructs are correlated with outcome constructs. As the correlation of the various constructs are positive and statistically significant (Table 2), criterion-related validity exists for these constructs.

SEM produces parameter estimates of links between the latent variables and so is also called latent variable analysis, or causal modeling. AMOS 4.0 and SPSS 10.1 (Arbuckle & Wothke, 1999) were employed for the SEM analysis.

### RESULTS AND DISCUSSION

In summary, this research examined the impact of systems’ quality, information quality, and information presentation on user satisfaction and intention to use in the context of choiceboard systems. The IS success model was used as the basis of the research model. The model was based on Shannon and Weaver’s (1949) communication theory, Mason’s (1978) theory, and the Delone and McLean (1992) model. The research model employed the constructs at the technical level (i.e., systems’ quality and information quality) in the context of choiceboards and finally its impact on different components of user satisfaction such as interface satisfaction, decision-making satisfaction, and resultant overall satisfaction. The path coefficients calculated for the estimated model support the hypothesized relationships in both direction and magnitude with few exceptions. Overall, the statistical conclusions support the research model (Figure 1).

System quality is directly and positively correlated to interface satisfaction (H-1); so, an increase in the quality of the system leads to an increase in satisfaction in using the interface. Information quality is directly and positively correlated to interface satisfaction (H-3); so, an increase in the quality of the information leads to an increase in satisfaction in using the interface. Information presentation is not directly and positively correlated to interface satisfaction; (H-5); therefore, this hypothesis is not validated.

The path coefficients calculated for the estimated model also support the hypothesized relationships in both direction and magnitude in the case of decision-making satisfaction. Most

<table>
<thead>
<tr>
<th>No.</th>
<th>Construct</th>
<th>Interface Satisfaction</th>
<th>Decision-making Satisfaction</th>
<th>Overall Satisfaction</th>
<th>Intention to Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>System Quality</td>
<td>0.66**</td>
<td>0.65**</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Information Quality</td>
<td>0.54**</td>
<td>0.69**</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Information Presentation</td>
<td>0.50**</td>
<td>0.44**</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Interface Satisfaction</td>
<td>-</td>
<td>-</td>
<td>0.49**</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>Decision-Making Satisfaction</td>
<td>-</td>
<td>-</td>
<td>0.51**</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>Overall Satisfaction</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.56**</td>
</tr>
</tbody>
</table>

**p < 0.01
of the hypotheses in the area of decision-making satisfaction have been validated using the data. System quality is directly and positively correlated to decision-making satisfaction (H-2); so, an increase in the quality of the system leads to an increase in decision-making satisfaction. Information quality is directly and positively correlated to decision-making satisfaction (H-3); so, an increase in the quality of the information leads to an increase in decision making. Presentation is not directly and positively correlated to decision-making satisfaction (H-6), as this hypothesis is not validated.

System quality includes system ease of use, convenience of access, and system reliability. Thus, a net positive effect from these factors will result in a positive effect on interface satisfaction and decision-making satisfaction. In choiceboards, as in other systems, the ease of use of the system, convenience of access, and system reliability are important considerations for the user. Information relevance, accuracy, completeness, and timeliness constitute the construct information quality. Thus, a net positive effect from these factors will result in a positive effect on decision-making quality. Choiceboard systems should provide relevant, accurate, complete, and timely information for better decision-making satisfaction.

Graphics, color, presentation style, and navigational efficiency measures information presentation. Therefore, information presentation measures how information is displayed. It was hypothesized that a net positive effect from graphics, color, presentation style, and navigational efficiency would result in a positive effect on interface satisfaction and decision-making satisfaction. The data did not support this hypothesis.

The statistical conclusions support the hypotheses on user satisfaction. Interface satisfaction is directly and positively correlated to overall satisfaction (H-7); so, an increase in interface satisfaction leads to an increase in overall satisfaction. Similarly, decision-making satisfaction is directly and positively correlated to overall satisfaction (H-8); so, an increase in decision-making satisfaction leads to an increase in overall satisfaction. Overall satisfaction also is found to be directly and positively correlated to intention to use (H-9); so, an increase in overall satisfaction leads to an increase in intention to use. The results from the research model also demonstrate the relative weight of system quality compared to information quality. Interestingly, decision-making satisfaction of end users, the quality of the system, is more important than the quality of the information.

As with all regression and structural equation modeling techniques, correlation does not prove the causality of the relation. However, since these causal relationships are based on an established literature, and since the theoretical grounding of the causality is adequate, it is reasonable to concur with the causality, where it has been validated (Gefen, Straub, & Boudreau, 2000).

MANAGERIAL IMPLICATIONS AND FUTURE RESEARCH

The research results empirically demonstrate the relationships between interface satisfaction, decision-making satisfaction, system quality, information quality, and information presentation. It also demonstrates the relationships among variables such as interface satisfaction, decision-making satisfaction, overall satisfaction, and the intention to use. These relationships are useful in influencing the intention to use among users of choiceboard systems. IS professionals need to understand these relationships to help their firms design choiceboard systems that are effective. This research provides an understanding of those interrelationships.

In the context of choiceboards, the quality of information influences decision-making satisfaction. For example, for a choiceboard system that allows users to develop their own holiday itinerary, the research suggests that users would value
Product Customization on the Web

complete, accurate, and relevant information about holiday sites, weather, local costs, flights, rentals, and hotels. Similarly, users will have better decision-making satisfaction with timely, accurate, and complete information as they develop alternative scenarios for their holidays.

The research suggests that ease of use, convenience of access, and system reliability also influence the decision-making satisfaction of users. A choiceboard, other than just being available and accessible, also should be easy to use. A user should not feel overwhelmed by available choices. The research also suggests that ease of use, convenience of access, and system reliability and flexibility influence interface satisfaction. The quality of the choiceboard system makes an impact if it is user-friendly, clear, and understandable. Interface and decision-making satisfaction influences if the choiceboard has been satisfactory in meeting user needs, which affects intention to use. For choiceboard users, it is not only important that the quality of the choiceboard system and the information it provides be adequate, but also that it provides them with interface and decision-making satisfaction. Thus, they will intend to use the choiceboard if they find it useful and if it meets their needs. This research shows that choiceboard users are deriving satisfaction with the system in a more complex fashion. If the choiceboard provides interface as well as decision-making satisfaction such that there is overall satisfaction, only then will they be a repeat user.

The empirical data suggest that the presentation of information is not important to the user in decision making. Users are not particularly impressed by color, graphics, and presentation style; they are more interested in the pertinent information being provided to them via the system. This is an interesting result, because in the recent past, there has been an increase in color and graphics on Web sites; but this presentation is of limited use if these Web sites are not able to provide the desired quality of information.

This research has examined the perceptions of users relative to their intention to use and how that perception is affected by overall satisfaction, which, in turn, depends on decision-making satisfaction and interface satisfaction. Much of the model has been validated by the data. Even the hypotheses that were not validated provide interesting insights. Studies should be conducted using other Web-based systems to test if the results of the present study can be extended to other situations. Qualitative studies also can be conducted to study choiceboard systems. These studies have the possibility of providing insight about choiceboard system users. These studies will help to build a wider body of research, which is needed for designing effective choiceboard systems.

REFERENCES


**ENDNOTE**

1 A preliminary version of the paper was published in the *Proceedings of 2004 Americas Conference on Information Systems.*

*This work was previously published in the Information Resources Management Journal, edited by M. Khosrov-Pour, Volume 19, Issue 2, pp. 69-81, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 5.20

Distributed Deception: An Investigation of the Effectiveness of Deceptive Communication in a Computer-Mediated Environment

Randall J. Boyle
University of Utah, USA

Charles J. Kacmar
The University of Alabama, USA

Joey F. George
Florida State University, USA

ABSTRACT

This research examines the impact of computer-mediated communication, distributed communication, and knowledge of prior baseline behavior on an individual’s propensity to make veracity judgments. This study found that more detection confidence can come from knowledge of a person’s prior baseline behavior, being proximally located, the type of communication media used, and perceived relational closeness. These factors indirectly lead to less deception detection through more detection confidence and reliance on the truth bias, a fundamental belief in the truthfulness of others, even in a computer-mediated environment.

INTRODUCTION

Today’s firms are relying more on non-traditional communication media such as e-mail, voicemail, and virtual meetings (Burke and Chidambaram, 1999; Guicking, Tandler and Grasse, 2008; Hoffman and Novak, 1996). In a globalized work environment, these types of communication media are important to quality decision-making (Fjermestad, 2005). However, a question arises...
as to whether the use of these communication media, together with distributed information sources, may mitigate an individual’s ability to detect deception. Deception is a regular part of daily communicative interaction (DePaulo and Kashy, 1998), accounting for 26-33% of daily social interactions (DePaulo, Kashy, Kirkendol, Wyer and Epstein, 1996; Hancock, Thom-Santelli and Ritchie, 2004). As computer-mediated communication use continues to spread, the ability to detect deception using lean communication media will be increasingly important in the workplace.

Research in the area of deception detection over distributed media has not been widespread. While individual research streams such as media richness, computer-mediated communication, and deception detection have extensive bodies of literature, the intersection of these streams has scarcely been examined (George and Marett, 2005). For example, deception detection research has focused on techniques such as training to recognize deceptive cues (Feeley and Young, 1998; Ekman and O’Sullivan, 1991), and suspicion arousal (Stiff and Miller, 1986) to increase face-to-face deception detection rates. Face-to-face deceptive cues such as greater pupil dilation, more blinking, decreased response length, more speech errors and hesitations, greater voice pitch, more negative statements, and more irrelevant information (Feeley and Young, 1998) are of limited applicability to deception detection in computer-mediated or distributed environments.

Similarly, the ability to detect deception between communication partners is widely believed to be related to the type of personal relationship that exists between them (Feeley and Young, 1998). Past research in the area of deception detection has mainly focused on individuals who are either strangers or intimate partners (Anderson, Ansfield, and DePaulo, 1997). However, working relationships typified by high levels of familiarity but low levels of intimacy, have been largely ignored. The relationship between partners impacts deception detection rates because of a fundamental assumption that their partner is being truthful. This fundamental assumption of truthfulness is often referred to as the truth bias (McCornack and Parks, 1986).

The purpose of this study is to investigate the effects of differing contextual factors on deception detection confidence and the relationship between confidence and truth bias. More specifically, this study will contribute to existing IS literature by examining the effects of working relationships, computer-mediated communication, and distributed environments within the context of deception detection. The next section of the paper presents the theoretical background for the study, including a research model and hypotheses. This is followed by a discussion of the research method, findings, and implications for research and practice.

THEORETICAL BACKGROUND

Research surrounding deception detection has focused on detection skills of observers (Brandt, Miller, and Hocking, 1982), conversational task demands (Burgoon and Newton, 1991), honesty judgments (Fiedler and Walka, 1993), the influence of relational closeness (Anderson et al., 1997), environmental influence (Storms, 1973), observer ability to detect deception (Buller, Strzyzewski, and Hunsaker, 1991), and the impact of suspicion on detection accuracy (Buller, Strzyzewski, and Comstock, 1991). Findings have shown that individuals have significant difficulty discerning truth from deception. Deception detection rates have been shown to range from 55% to 60% (Feeley and Young, 1998). However, other studies have found that deception detection rates may be as low as 35-40%, while truth detection rates have ranged from 70-80% (Levine, McCornack, and Park, 1998). It is important to note that deception detection and truth detection, correctly identifying lies as lies and truths as truths respectively, vary in overall task difficulty.
Deception Models

In order to identify the factors influencing differences in deception detection, several theoretical models were developed and later empirically tested. McCornack and Parks (1986) developed an important theoretical model looking at the antecedents of the truth bias. Specifically, they looked at the impact of relational closeness on deception detection. They found that as relationships develop and people become closer, they tend to be more confident of their ability to detect truths/lies. This increased confidence causes participants to subsequently rely more on their basic assumption of truthfulness (i.e., the truth bias).

The model, shown in Figure 1, has received additional support in a recent meta-analysis indicating a strong linkage between relational closeness and detection confidence (DePaulo, Charlton, Cooper, Lindsay, and Muhlenbruck, 1997). As detection confidence increases, individuals will reduce the amount of effort they put forth in detecting deceptive behavior by relying increasingly on truth bias and reducing detection accuracy. Some scholars have operationalized truth bias as the percentage of truthful judgments during lie detection; others as the number of truthful judgments (McCornack and Parks, 1986); and yet others have operationalized it using self-report measures (Stiff, Kim, and Ramesh, 1992).

Table 1. Concept definitions

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth Bias</td>
<td>The inherent belief that people are telling the truth (Stiff, Kim, and Ramesh, 1992)</td>
</tr>
<tr>
<td>Relational Closeness</td>
<td>The degree of interdependence that exists between relationship partners as evidenced by their day to day activities (Berscheid, Snyder, and Omoto, 1989)</td>
</tr>
<tr>
<td>Baseline Knowledge</td>
<td>Knowledge of a person’s nominal/truthful behavior (Feeley, deTurck, and Young, 1995)</td>
</tr>
<tr>
<td>Proximity</td>
<td>The physical presence or absence of individuals in the same room</td>
</tr>
<tr>
<td>Deception Detection</td>
<td>The ability of an individual to recognize deviations from nominal, or truthful, behavior (Miller, Mongeau, and Sleight, 1986)</td>
</tr>
<tr>
<td>Detection Confidence</td>
<td>Belief that a truth or lie is correctly identified</td>
</tr>
<tr>
<td>Detection Accuracy</td>
<td>Correctly identifying truths as truths, and lies as lies (Feeley and Young, 1998)</td>
</tr>
<tr>
<td>Deception</td>
<td>A message knowingly transmitted by a sender to foster a false belief or conclusion by the receiver (Knapp and Comadena, 1979)</td>
</tr>
<tr>
<td>Deception Accuracy</td>
<td>Correctly identifying lies as lies (Feeley and Young, 1998)</td>
</tr>
<tr>
<td>False Accusation</td>
<td>Incorrectly identifying a truthful statement as a lie.</td>
</tr>
<tr>
<td>Information Cue</td>
<td>Acts that give away information the sender wishes to conceal (Zuckerman and Driver, 1985)</td>
</tr>
</tbody>
</table>
Stiff, et al. (1992) similarly showed that there is a significant relationship between relational development and truth bias. They note that there is substantial support for the claim that more developed relationships, typified by intimacy, reduced uncertainty, satisfaction, and commitment, and tend to be much more subject to truth bias. They also note that participants in those relationships are only marginally less able to detect deception. In turn, participants who rely more on the truth bias in making veracity judgments also tend to view their partner’s statements as truthful. Stiff, et al. (1992) found that aroused suspicion had significant negative correlations with both truth bias and judgments of truthfulness. However, the study procedures used to induce heightened suspicion may produce findings that are only generalizable to situations where individuals already anticipate deception (Miller and Stiff, 1993). In sum, both theoretical models suggest that an increase in relational closeness will increase the level of truth bias and reduce deception detection.

**RESEARCH MODEL**

The research model shown in Figure 2 depicts the new constructs and relationships tested in this study. The model hypothesizes associations from baseline knowledge, detection confidence, and truth bias to deception detection. The model extends the work of McCornack and Parks (1986).
with constructs for baseline knowledge, media richness, and proximity. Unlike prior studies, the research model presented distinguishes baseline knowledge from relational closeness in order to separately test the influences of relational intimacy and knowledge of prior baseline behavior on detection confidence. Prior research has also focused on deception detection within nominal groups. This study adds to existing theory by examining the impact of computer-mediation and distributed communication on deception detection. These new constructs will allow us to investigate the impact of new methods of communication on deception detection.

In deception detection research, it is important to note the distinction between deception detection and detection accuracy. Deception detection refers to the correct identification of a deceptive statement, while detection accuracy refers to the correct identification of a truthful statement, a deceptive statement, or both. In order to distinguish deception detection from a broader definition of detection accuracy, this study measures deception detection, truth detection, successful deception, and false accusations separately. Figure 2 shows the decomposition of the construct detection accuracy into four separate dependent variables. Decomposing detection accuracy into four separate variables provides a more rigorous methodology and greater clarity when interpreting outcomes.

**Relational Closeness and Baseline Behavior**

As previously mentioned, baseline behavior is a key factor in predicting deception detection. In order for an individual to recognize deviations from nominal or truthful behavior he or she must first acquire a “baseline” of truthful behavior. This baseline knowledge is important because an individual may naturally exhibit cues typically associated with deceptive behavior (Brandt, et al., 1982). Feeley and Young (1998) note that the majority of deception detection studies are based on the premise that there is a set of identifiable cues that are specifically related to deceptive behavior. However, other researchers argue that identifying changes in individual-specific behaviors during deception greatly reduces the possibility of misinterpretation (Anderson, et al., 1997; Walters, 2000). Brandt, Miller, and Hocking (1980) found that observers with truthful baseline knowledge were three times as accurate in detecting deceptive behavior as individuals without truthful baseline knowledge. Similar studies have shown that observers with a small amount of baseline knowledge had better detection rates (72%) than observers without baseline knowledge (56%) (Feeley, deTurck, and Young, 1995). The fact that many previous studies have used zero-history dyads may well explain the low detection rates. Constructs such as suspicion (Stiff and Miller, 1986), anonymity (Kahai and Avolio, 2006), and trust (McKnight, Cummings, and Chervany, 1998) have also been shown to influence group interactions. However, the addition of these constructs to the research model is beyond the scope of this paper.

While baseline knowledge is vital in making veracity judgments, more relational closeness, which often comes in obtaining baseline knowledge, can severely mitigate any advantage. Miller and Stiff (1993) demonstrated that individuals in intimate relationships may not want to view their partners as deceptive due to emotional attachments. In fact, studies have shown that relational intimacy is unrelated to detection accuracy (Comadena, 1982). McCornack and Parks (1986) examined the relationship between relational closeness, truth bias, and detection confidence and found that increased relational closeness decreases the ability to detect deceptive behaviors. This leads to the conclusion that, while knowing truthful baseline behaviors and cues are essential to deception detection, relational closeness is detrimental to veracity judgments and significantly increases an individual’s level of truth bias.
Knowledge of the other person’s baseline behavior is predicted to have a positive influence on detection confidence and detection accuracy. In other words, as individuals become more able to “read” their communication partner, their ability to detect deceptive behavior will increase. Detection accuracy can be broken down into deception detection accuracy and truth detection accuracy. Knowledge of baseline behavior is predicted to have a positive influence on truth detection and deception detection. Individuals will be able to detect truthful cues, deceptive cues, and deviations from nominal behavior. With regards to social interaction, the model predicts that optimal deception detection is obtained by minimizing relational closeness, not being proximally located, and maximizing the amount of baseline behavior knowledge. By minimizing intimacy and maintaining familiarity in a relationship, an individual should theoretically be able to detect deception. Thus, the following hypotheses are put forth:

**Hypothesis 1:** Knowledge of baseline behaviors will be positively associated with deception detection.

**Hypothesis 2:** Knowledge of baseline behaviors will be positively associated with detection confidence.

**Hypothesis 3:** Relational closeness will be positively associated with detection confidence.

**Communication Media**

As previously noted, media richness theory (MRT) suggests that as media become richer, the amount of information cues conveyed increases (Daft and Lengel, 1986). Within the context of truth bias and deception detection, the ability to detect information cues is critical in evaluating the truthfulness of a message (Zuckerman and Driver, 1985). Information cues can consist of verbal, visual, paralingual, and textual cues (Williams, 1977; Zuckerman, DePaulo, and Rosenthal, 1981). Cues provide receivers with the evidence necessary to make veracity judgments. It has been shown that individuals assume a person is telling the truth until additional cues indicate otherwise (McCornack and Parks, 1986). For this study, deception will be defined as a message knowingly transmitted by a sender to foster a false belief or conclusion by the receiver (Knapp and Comadena, 1979).

Zuckerman and Driver’s (1985) meta-analysis of 24 possible deceptive behaviors yielded 14 specific behaviors that were related to deception. An analysis by Feeley and Young (1998) shows that eight cues are significantly correlated with deception: greater pupil dilation, more blinking, decreased response length, more speech errors and hesitations, greater voice pitch, more negative statements, and more irrelevant information. Recent studies have explored the ability to detect deception using linguistic based cues (Zhou, Burgoon, Nunamaker, and Twitchell, 2004).

The ability for decision makers to make correct veracity judgments may be dependent on the quantity, quality, and variety of information cues (McCornack, 1997). The quantity, quality, and variety of information cues is dependent upon the type of media used. Media selection may be, as suggested by Media Richness Theory, dictated by the need to make veracity judgments. Media Richness Theory (MRT) (Daft and Lengel, 1986) has been used to explain the effects and choice of communication media in a variety of tasks such as planning, creativity, intellectual tasks, decision making, and brainstorming (Hollingshead, McGrath, and O’Connor, 1993). Prior research has also shown that partners with a history of working together communicate differently across varying types of media (Yoo and Alavi, 2001; Carlson and Zmud, 1999). Media richness can be defined as the ability of information to change understanding within a given time interval (Daft and Lengel, 1986; Dennis and Kinney, 1998). Depending on the degree of equivocality and task uncertainty,
an individual will choose a richer medium in order to provide better feedback, draw upon a wider variety of cues, use a variety of language, and convey a more personally focused message (Daft, Lengel, and Trevino, 1987). With respect to the current study, the focus is not on what type of media is best for deception detection, but rather, how different levels of media richness affect an individual’s ability to make veracity judgments and mitigate the truth bias.

Communication media, or the richness of the communication media, is predicted to have a positive influence on detection confidence, and detection accuracy. Increased media richness increases the number of information cues individuals can use to make veracity judgments. More cues should lead to higher levels of confidence in deception detection because of reduced equivocality. Thus, the following hypotheses will be tested in this study:

**Hypothesis 4:** The richness of the communication medium will be positively associated with deception detection.

**Hypothesis 5:** The richness of the communication medium will be positively associated with detection confidence.

**Proximity**

Social facilitation, as conceptualized by Zajone (1965), focuses on changes in behavior due to the physical presence of others. In a review of 287 studies examining the effects of social facilitation, Guerin (1986) notes that the presence of another person may cause behavioral changes in four ways. First, the physical presence of another individual may cause more apprehension due to the expectation that he or she will be evaluated. Second, the physical presence of someone else may cause “cognitive or physical conflict” between task focus and having to attend to the other person. Third, the presence of another person may enhance task performance in order to make a good impression on the other person. Fourth, another’s physical presence may cause an increase in “conforming to public and private norms” because of greater self-focus and comparison of behavioral standards (Guerin, 1986).

Distributed communication has been shown to have a significant impact on the final outcome of a process, communication act itself (Rice, 1984), and the effort an individual puts into maintaining social networks (Miranda and Carter, 2005). Research has also shown that distributed teams can experience escalation in conflict (Armstrong and Cole, 1995) and the formation of hostile coalitions (Cramton, 2001). Whitty (2002) showed that there was a propensity for anti-social behaviors in chat rooms where individuals were relatively new. However, more intimate relationships did develop over time. Despite the significant effects of presence on behavior, the influence of proximity and computer mediation has typically been difficult to separate due to methodological inconsistencies (Hedlund, Ilgen, and Hollenbeck, 1998).

In sum, distributed communication decreases the level of social influence, thereby changing individual behavior and expectations. The presence of another person encourages an individual to conform to social norms and to restrain from engaging in anti-social behaviors (Berger, Hampton, Carli, Grandmaison, Sadow, and Donath, 1981). Proximally located subjects may be more confident of their ability to detect deception due to their belief that their communication partner is more likely to conform to social norms (i.e., be truthful).

The proximal influence of another person has not only been shown to cause a change in behavior, but also a change in the level of cognitive effort put forth (Geen and Gange, 1977; Zajonc, 1965). Individuals will put forth more effort toward task performance in the presence of others. Subjects may put forth more effort in detecting deception when proximally located to their communication partner. As a result, subjects may be more
confident in their ability to detect a deceptive decision. In this study it is predicted that physical proximity will influence individuals to be more confident in detecting deception. Therefore, the following hypothesis is put forth:

**Hypothesis 6:** Proximally located subjects will have higher detection confidence than subjects that are distributed.

**Truth Bias and Deception Detection**

Truth bias is the inherent belief that people are telling the truth (Stiff, Kim, and Ramesh, 1992). While this fundamental assumption of truthfulness simplifies decision making by reducing the quantity and quality of veracity judgments, it may also inhibit deception detection. Research on truth bias has been aimed at trying to reduce an individual’s reliance on this fundamental assumption of truthfulness. Studies have explored reduction techniques such as training (Ekman and O’Sullivan, 1991) and suspicion arousal (Stiff and Miller, 1986). In these studies, subjects were trained to recognize cues (e.g., eye twitching, specific words) thought to be specifically associated with deception, or were told that their communication partner may be lying. Efforts have also been made to reduce truth bias by lowering the level of cognitive demand through passively observing dialogue (Buller, et al., 1991). This was accomplished by allowing subjects to watch video recordings or read the dialogue between two other subjects. In general, research has focused on enabling individuals to better detect deception.

As hypothesized and reported by McCornack and Parks (1986), the level of confidence an individual has in his or her assessment of truthfulness should positively affect his or her reliance on the truth bias as a simplifying heuristic. In turn, greater dependence on the truth bias should negatively affect his or her ability to detect deception and lead to fewer false accusations due to the general belief that his or her communication partner is telling the truth. Conversely, the truth bias should have a positive affect on an individual’s ability to detect truthful statements and the individual’s susceptibility to being successfully deceived. The authors argue that as individuals become more certain of their evaluation, they devote fewer cognitive resources toward determining truthfulness. Subsequently, they will rely more on cognitive heuristics to make veracity judgments and, therefore, will be less able to distinguish between truthful and deceptive statements. Thus:

**Hypothesis 7:** Detection confidence will be positively associated with truth bias.

**Hypothesis 8:** Truth bias will be negatively associated with deception detection.

**Hypothesis 9:** Truth bias will be positively associated with truth detection.

**Hypothesis 10:** Truth bias will be positively associated with successful deception.

**Hypothesis 11:** Truth bias will be negatively associated with false accusation.

**METHOD**

This study utilized a controlled laboratory experiment employing a 2x2x2 factorial design (computer-mediated/non-computer-mediated, proximal/distributed, familiar/unfamiliar with communication partner). A total of 97 dyads were randomly assigned to one of eight possible treatment conditions (N=194). Dyads comprised of subjects familiar with each other were recruited from groups in the preceding semester taught by the first author. Proximal dyads completed the task in the same room, while distributed dyads completed the task in separate rooms (even though the rooms were adjacent to each other, subjects
could not see each other or communicate directly). Computer-mediated dyads exchanged messages using the Microsoft NetMeeting® chat software. Non-computer-mediated groups communicated using either face-to-face or cellular telephone.

**Task**

The prisoner’s dilemma task, developed by Albert Tucker (c.f., Axelrod, 1984), was chosen because of its characteristics in studying deception detection. The classic version of the prisoner’s dilemma task does not allow for communication between subjects. However, this study used a derivative of the prisoner’s dilemma task that allowed for real-time interpersonal interactions where subjects were motivated to deceive and detect deception. Anticipation of deception was not heightened by the researchers, nor was deception sanctioned in the experiment. Scoring allowed for measurement of deception, deception detection, and truth bias. Matching of participants with prior experience allowed individuals to use their prior knowledge to detect deception. Lastly, playing the prisoner’s dilemma game does not raise any serious ethical or human subjects concerns.

In the classical prisoner’s dilemma game, there are two players. Players can choose to either “defect” or “collude.” Each player makes his or her decision without the knowledge of what the other person will do. Defection always yields a higher payoff than collusion. The dilemma is that if both players defect, they will loose more than if they had both colluded. In other words, the greatest positive and negative payoffs are achieved when one player defects and his or her partner colludes, respectively. Since both players are trying to obtain this optimal solution, it is in their best interests to convince each other that they are going to collude and then defect.

Within the context of the game subjects were given the options to either “confess” (defect) or “stay quiet” (collude). Possible decision outcomes have the following reward structure \( T > C > 0 > D > S \) (Dawes and Orbell, 1995). The “temptation” payoff \( T \) of +$5 was rewarded for unilateral defection of one player, while the other remains silent. The “cooperative” payoff \( C \) of +$1 was rewarded for joint cooperation between individuals (i.e., both remain silent). The “defect” payoff \( D \) of -$1 was rewarded for joint defection of both individuals (i.e., both confess). The (more negative) “sucker” payoff \( S \) of -$3 was rewarded for unilateral cooperation. Negative dollar payoffs were given for mutual defection, and even greater negative dollar payoffs for unilateral cooperation.

*Figure 3. Detection matrix*
Distributed Deception

Figure 3 shows how each combination of decisions form the final dependent variables in the purposed research model.

Consistent with prior research, participants were able to lose all of their money, or more than double the amount they possessed before the game began. Subjects were told that they were to be given $3 for participating in the experiment and would gain or lose money according to their decisions. Researchers paid each subject in cash a week after the conclusion of the experiment. Subjects were paid at a later time to avoid any possible reprisal from another subject due to defection during the experiment.

Measures

The constructs in the research model were operationalized as described below. Relational closeness was operationalized using a 14-item short form of the relational closeness inventory (RCI) 27-item measure from Berscheid, Snyder, and Omoto (1989) (Polimeni, Hardie, and Buzwell, 2002). These items assess the strength of the relationship and frequency of contact between conversational partners. Items are measured using seven-point Likert scales with higher scores indicating a higher level of relational closeness. The RCI scale has been shown to be highly consistent, with a coefficient alpha of .90 (short form Cronbach alpha = .87) (Polimeni, et al., 2002).

Baseline knowledge was operationalized by pre-assigning subjects to the baseline knowledge condition if they had worked together on a large group project for three months during the semester prior to the experiment. Subjects who had never met before were assigned to the no baseline knowledge condition. Subjects were coded 2 = baseline knowledge and 1 = no baseline knowledge depending on their respective experiment condition. To ensure that subjects were correctly identified as having sufficient baseline knowledge a manipulation check consisting of three previously validated measures (Cronbach alpha = .90) was implemented. These measures indicated a subject’s general knowledge about their communication partner, nature of their relationship, and length of time they have known each other (Berscheid, et al., 1989).

Proximity, an individual’s awareness that another person is proximally located, was indicated as a factor of the treatment condition. Proximal subjects were seated such that they faced each other at a distance of no greater than three feet, while subjects in the distributed condition were seated in separate rooms, and could not see or communicate visually or verbally with each other. Responses were coded 2 = proximal and 1 = distributed, depending on their respective condition.

Richness of the communication media was operationalized using eight seven-point items (Cronbach alpha = .85) (Daft and Lengel, 1986; Dennis and Kinney, 1998). In the CMC group, subjects communicated with their partner through a chat program. In the non-CMC group subjects communicated face-to-face or by cellular telephone.

Detection confidence was operationalized using a single item that asked subjects their level of confidence in predicting their communication partner’s choice (confess, stay quiet). This self-report measure ranged from 0 (not sure) to 100 (very sure) with a mean of 86.13 (S.E. = 18.512).

Truth bias was operationalized using four seven-point Likert measures of perceived truth bias (Stiff, Kim, and Ramesh, 1992). Truth bias was computed as an average with scores ranging from 1 to 7 and a mean of 5.6 (S.E. = 1.18).

The four constructs: truth detection, deception detection, successful deception, and false accusation, were calculated based on subjects’ decisions to either “stay quiet” or to “confess”. Deception detection was operationalized as a correct estimation of a communication partner’s choice to confess. Truth detection was operationalized as a correct estimation of a communication partner’s choice to stay quiet. Successful deception and false accusation were operationalized as incorrect
estimations of a communication partner’s choice to confess or stay quiet.

SAMPLE

Subjects (N=194) were juniors and seniors majoring in management information systems (MIS) at a large southeastern university in the U.S. Subjects’ mean age was 23.4 years (S.E.=4.3) with 58.8% of the subjects being male. To ensure that subjects were correctly identified as having prior baseline knowledge of their partner’s behavior, subjects were asked how long they had known their partner. Subjects in the baseline group indicated that they had known their partner for an average of 9.18 months (S.E.=8.5) and had a mean relational closeness measure of 1.72 (S.E.=0.094). Individuals in the no-baseline condition indicated that they had known their partner for an average of 0.85 months and had a mean relational closeness measure of 1.29 months (S.E.=0.074). The mean difference in the time subjects had known their partner in the baseline and no-baseline groups was statistically significant (F=13.172; p<0.001).

A manipulation check asking subjects if they could see their partner indicated that the proximal/distributed manipulation was successful and statistically significant (F=45.335; p<0.001). To address a possible concern of information leakage in proximal groups, subjects were asked if they were able to gain information from overhearing other conversations during the experiment. Responses across all conditions indicated that subjects were not able to gain information from other groups (Mean = 1.28; S.E. = 0.737).

ANALYSIS AND RESULTS

Scatter-plots and histograms indicated that some measures experienced minor departures from normality. The distribution of detection confidence scores was somewhat peaked (Kurtosis=2.06). The distribution for relational closeness was also peaked (Kurtosis=4.02) and somewhat skewed (Skewness=2.09). Despite minor departures from normality, the F statistics reported in this study were robust. Kirk (1995) notes that the F statistic remains robust when “populations are symmetrical, but not normal, and the sample sizes are equal or greater than 12.” (Kirk, 1995; Clinch and Keselman, 1982). Analysis of histograms across experiment conditions indicated that populations were homogeneous in form (positively skewed), and sample sizes were approximately equal. Therefore, minor departures from normality were not a threat to the validity of the statistical procedures used in this study.

Factor Analysis

Three of the seven constructs (relational closeness, truth bias, and communication media) were measured using multiple items to identify specific underlying latent constructs in the measurement model. The remaining independent, intermediate, and dependent variables were operationalized using single measures. Factor analysis was used to condense information gathered from multiple variables into a minimal number of factors representing the theoretically proposed latent constructs. The dimensionality of relational closeness (RC), truth bias (TB), and communication media (CM) were determined using principal components analysis (PCA) (Kaiser-Meyer-Olkin=0.796).

A series of initial unconstrained principal component analyses were done to determine the amount of variance explained by the latent factors. Initial analyses indicated that several items did not explain a significant amount of variance. Factors with loadings below .50 or cross-loadings above .30 were dropped (Hair, Anderson, Tatham, and Black, 1998). Factors having loadings of .50 or greater are considered practically significant and a factor will account for 50% of the variance if the factor loading exceeds .70 (Hair, et al., 1998). Based on these criteria, seven of the items
in the relational closeness factor and two items from the media richness factor were eliminated. All other constructs yielded acceptable statistics and, therefore, no other items were dropped. The follow-up factor analysis revealed that all factors had loadings of .70 or greater except for RC2 (.681) when using a constrained three-factor solution. There were no cross loadings greater than .25.

Table 2 shows the variance explained by each item and respective factor loadings. Varimax rotation was used throughout the factor analysis to give a better interpretation of the underlying factors by reducing the ambiguities that often accompany initial unrotated factor solutions (Hair, et al., 1998).

All factors exhibited reliabilities consistent with prior research scoring above .70. Constructs exhibited convergent and discriminant validity by having Cronbach’s alpha values greater than 0.70 (Nunnally, 1978) and by each indicator loading more highly on its associated construct than any other construct. Means, Standard Errors, Cronbach’s alphas and intercorrelations for all constructs can be seen in Table 3.
Distributed Deception

Research Model and Hypothesis

Results

Partial Least Squares (PLS) was used to test the research model because it accommodates limited sample sizes, a large number of constructs, and restrictions on measurement scales (Chin, Marcolin, and Newsted, 2003). PLS has also been shown to be robust when dealing with discrete single-indicator variables (Pavlou and Fygenson, 2006). While PLS is typically used in exploratory research, it can also be used in confirmatory research if sample size is an issue (Gefen, Straub, and Boudreau, 2000). Given the relatively small sample size used in this study, PLS is considered an appropriate modeling technique. Path coefficients and $R^2$ values for the purposed research model are shown in Figure 4.

Hypothesis 1 (H1) stated that knowledge of baseline behaviors will be positively associated with deception detection. Surprisingly, individuals with baseline knowledge about their partner had a significantly lower deception detection accuracy rate ($b=.063$, $t=2.06$, $p=0.020$) than those without. Overall, deception detection rates were 31.6% for baseline groups and 68.4% for groups with no baseline knowledge. A “hit” included successful detection of deceptive decisions. Thus,

| Relational Closeness (RC) | 1.50 | 0.860 | 0.86 |
| Baseline Knowledge (BK) | 1.53 | 0.501 | 0.27 | *
| Proximity (P) | 1.47 | 0.501 | -0.10 | 0.01 | *
| Communication Media (CM) | 4.89 | 1.31 | 0.00 | 0.14 | 0.12 | 0.86 |
| Detection Confidence (DC) | 86.1 | 18.4 | 0.13 | 0.28 | 0.13 | 0.19 | **
| Truth Bias (TB) | 5.61 | 1.17 | 0.02 | 0.30 | 0.10 | 0.33 | 0.52 | 0.85 |
| Deception Detection (DD) | 0.979 | 0.214 | -0.12 | -0.10 | 0.00 | -0.05 | -0.04 | -0.15 | **
| Truth Detection (TD) | 0.717 | 0.324 | 0.08 | 0.19 | 0.11 | 0.07 | 0.17 | 0.36 | -0.52 | **
| Successful Deception (SD) | 0.119 | 0.023 | -0.02 | -0.09 | -0.07 | 0.04 | -0.01 | 0.00 | -0.12 | -0.58 | **
| False Accusation (FA) | 0.067 | 0.018 | 0.01 | -0.09 | -0.12 | -0.12 | -0.24 | -0.49 | -0.09 | -0.43 | -0.09 | **

Table 3. Intercorrelations and reliabilities
we reject H1. However, we did find a significant relationship in the opposite direction.

Hypothesis 2 (H2) stated that knowledge of baseline behaviors will be positively associated with detection confidence. Detection confidence was scored on a scale from 1 (very unconfident) to 100 (very confident). Subjects in the baseline condition (N=92) had a mean detection confidence level of 91.54 (S.E.=15.9), while subjects in the no baseline condition (N=102) had a mean detection confidence level of 81.31 (S.E.=19.2). Individuals with baseline knowledge were significantly more confident (b=.226, t=3.32, p<0.001) than those without baseline knowledge. Thus, there is evidence to support H2.

Hypothesis 3 (H3) stated that more relational closeness will be positively associated with detection confidence. Consistent with prior research, results confirmed that relational closeness had a positive relationship on detection confidence (b=.120, t=2.56, p=0.006). Therefore, there is sufficient evidence to support H3.

Hypothesis 4 (H4) stated that the richness of the communication medium will be positively associated with deception detection. Table 4 shows deception detection rates, attempts at deception, successful deception rates, and deception detection confidence rates. Subjects in the non-CMC condition did not have significantly higher detection accuracy rate than subjects in the CMC condition (b=0.00, t=0.685, p<0.247). Overall, deception detection rates were low for both manual (42.9%) and computer (47.6%) groups where subjects were deceived. Thus, we find that there is insufficient support for H4.

Hypothesis 5 (H5) stated that the richness of the communication medium will be positively associated with detection confidence. Subjects in the non-CMC condition communicated with their partner either face-to-face or by cellular telephone, while subjects in the CMC condition communicated with their partner using NetMeeting®. Subjects in the manual condition (N=94) had a mean detection confidence level of 89.61 (S.E.=16.8), while subjects in the computer condition (N=100) had a mean detection confidence level of 82.92 (S.E.=19.4). Results indicate that the means were significantly different (b=.151, t=3.17, p<0.001). Thus, H5 was supported.

Hypothesis 6 (H6) stated that proximally located subjects will have higher detection confidence than subjects that are distributed. Subjects
in the proximal condition (N=102) had a mean detection confidence level of 88.45 (S.E.=16.8), while subjects in the distributed condition (N=92) had a mean detection confidence level of 83.62 (S.E.=19.8). Results showed that proximally located subjects did have significantly higher detection confidence than subjects that were distributed (b=.128, t=2.19, p=.015). Overall, baseline knowledge, communication media, proximity, and relational closeness accounted for 12.9% of the variance in detection confidence (R^2=.129). Thus, there is sufficient evidence to support H6.

Hypothesis 7 (H7) stated that detection confidence will be positively associated with truth bias. Results showed that detection confidence had a positive effect on truth bias (b=.519, t=8.75, p<0.001). Detection confidence accounted for 26.9% of the variance in truth bias (R^2=.269). Therefore, there is sufficient evidence to support H7.

Hypotheses 8 (H8) and 11 (H11) stated that truth bias will be negatively associated with detection and false accusation respectively. Truth bias had a negative effect on detection (b=-.137, t=1.97, p=.025) and a negative effect on false accusations (b=-.485, t=6.38, p<0.001). Truth bias and baseline knowledge accounted for 2.8% of the variance in detection detection (R^2=.028), and truth bias alone accounted for 23.5% of the variation in false accusations (R^2=.235). Therefore, H8 and H11 were supported.

Hypotheses 9 (H9) and 10 (H10) stated that truth bias will be positively associated with truth detection and successful deception, respectively. Truth bias had a positive effect on truth detection (b=.360, t=4.46, p<0.001) and a positive effect on successful deceptions (b=.016, t=1.37, p<0.086). Truth bias accounted for 13.0% of the variance in truth detection (R^2=.130), and truth bias accounted for less than 0.1% of the variation in false accusations (R^2=.001). Therefore, H9 and H10 were supported.

**DISCUSSION**

This study investigated the direct and indirect effects of differing contextual factors on deception.
Distributed Deception

detection and the relationship between detection confidence and truth bias. The contextual factors studied were the closeness of the relationship between the communicating partners, their baseline knowledge of each other, whether the partners were co-located or dispersed, and the particular media used for communication. All four hypotheses dealing with detection confidence were supported. Subjects who were familiar with their communication partners, who had baseline information about them, were co-located, and who used richer communication media had higher levels of confidence in their abilities to detect deception in their partners compared to subjects in the opposite conditions. The positive effects of baseline knowledge, richer communication media, and proximity on deception confidence are significant contributions of this study to existing research. High levels of detection confidence also were found to be associated with high levels of truth bias, supporting another hypothesis, and confirming that existing deception theories are robust within computer-mediated environments.

Of the two factors thought to directly affect deception detection (baseline knowledge, communication media), only the link from baseline knowledge to deception detection was supported. Further, this relationship was negative. This finding indicates that in some work situations, there will be little or no relationship between communication media and deception detection. Results confirm the existence and potency of the truth bias within computer-mediated environments by

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>t</th>
<th>p</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>2.0670</td>
<td>0.0200</td>
<td>Reject†</td>
</tr>
<tr>
<td>H2</td>
<td>3.3222</td>
<td>0.0005</td>
<td>Accept*</td>
</tr>
<tr>
<td>H3</td>
<td>2.5614</td>
<td>0.0056</td>
<td>Accept*</td>
</tr>
<tr>
<td>H4</td>
<td>0.6849</td>
<td>0.2471</td>
<td>Reject</td>
</tr>
<tr>
<td>H5</td>
<td>3.1669</td>
<td>0.0009</td>
<td>Accept*</td>
</tr>
<tr>
<td>H6</td>
<td>2.1990</td>
<td>0.0145</td>
<td>Accept*</td>
</tr>
<tr>
<td>H7</td>
<td>8.7516</td>
<td>0.0001</td>
<td>Accept*</td>
</tr>
<tr>
<td>H8</td>
<td>1.9730</td>
<td>0.0250</td>
<td>Accept*</td>
</tr>
<tr>
<td>H9</td>
<td>4.4594</td>
<td>0.0001</td>
<td>Accept*</td>
</tr>
<tr>
<td>H10</td>
<td>1.3724</td>
<td>0.0858</td>
<td>Accept**</td>
</tr>
<tr>
<td>H11</td>
<td>6.3836</td>
<td>0.0001</td>
<td>Accept*</td>
</tr>
</tbody>
</table>

* Significant at p< 0.05, ** Significant at p< 0.10, † Significant at p< 0.05 but in reverse direction.
negatively influencing deception detection and false accusations. Conversely, the truth bias was shown to positively influence truth detection and successful deception. Each of these findings is explored in more depth below.

Knowledge of Baseline Behavior

As noted earlier, a surprising finding of this study is that subjects with prior knowledge of their partner’s baseline behavior were less accurate in detecting deceptive decisions made by their communication partners. Past research has indicated that individuals who have known their partner for a short period of time should be more accurate in detecting deceptive behavior (Brandt, Miller, and Hocking, 1980). Past studies looking at deceptive behavior examined the ability of individuals to detect deception in intimate relationships, between friends, and among strangers (Anderson, et al., 1997). The proposed model theorized that familiarity with an individual’s baseline behaviors would aid in detecting deception cues. Results, however, indicate that even a small amount of familiarity reduces an individual’s ability to detect deception. This may be due to increased levels of trust and/or individuals’ inability to correctly recognize deceptive behaviors. Thus, by separating baseline knowledge from relational closeness, this study adds to existing theory by showing that both familiarity and intimacy (Feeley and Young, 1998; Comadena, 1982) are separate factors contributing to decreased deception detection.

Results indicate that more knowledge of baseline behaviors was significantly associated with more detection confidence. This suggests that individuals draw upon past experience when making veracity judgments and feel more confident about those judgments. Overall, the effect of having knowledge of a partner’s prior behavior caused individuals to be more confident in their veracity judgments, yet were significantly less accurate in those judgments.

Communication Media

This study extends existing deception detection theory by showing that individuals who communicate with their communication partner via face-to-face interaction or telephone are significantly more confident of their ability to detect deceptive behaviors, yet are not any better at actually detecting deception. In fact, they were indirectly less able to detect deception due to more detection confidence and subsequently more truth bias. Changes in behavior and decision making due to richer communication is supported by prior research (Barkhi, Jacob, and Pirkul, 1999).

In the context of deception detection, a richer communication media provides more information cues that might be used to detect deception. More truthful/deceptive cues in a richer communication medium can lead to more confidence in veracity judgments by providing more evidence and reducing equivocality. While cues may cause individuals to be more confident, it is the inherent uncertainty in discerning between truthful and deceptive statements that makes deception detection so difficult. In fact, most successful deceptions consist of many truthful statements and contain only a small, albeit critical, number of deceitful statements. By providing more informational cues that might indicate deceptive/truthful behavior, an individual may have greater justification for making a veracity judgment. Of course the results presented here are not meant to imply that information cues are not important. Rather, a minimal amount of information cues may be necessary to make a simple veracity judgment within the context of this study.

Proximity

In this study the physical presence of another individual led to more detection confidence due to the fact that individuals most likely felt compelled to conform to public and private norms.
In this study the public norm would be to tell the truth. Individuals in the proximal condition may have assumed that their communication partner would tell the truth to make a good impression and to avoid possible confrontation (Guerin, 1986) thus leading to more confidence in their decision. The converse would also be true in the distributed condition. As noted earlier, distributed communication decreases the level of social influence, thereby changing individual behavior and expectations. This study adds to existing theory by showing how proximity affects deception confidence, and ultimately deception detection, in computer-mediated environments. It should be noted that more detection confidence does not necessarily imply more deception detection.

**Relational Closeness**

As noted earlier, prior research co-mingled the constructs familiarity and relational closeness (Anderson, et al., 1997). This study has expanded existing research by showing that relational closeness itself is a contributing factor to detection confidence separate from baseline knowledge. This finding was significant (p=0.045) and consistent with prior research (DePaulo, et al., 1997).

Relational closeness is believed to have a direct, positive influence on detection confidence because, as feelings of closeness to another person grow, an individual will feel more confident of his or her ability to detect deception/truthfulness. Prior research indicates that this confidence may stem from the fact that individuals in intimate relationships may want to view their partners as truthful because of emotional attachments (Miller and Stiff, 1993). The results from this study imply that even minor levels of relational closeness developed over the course of a few months in work-related relationships might lead to significantly more detection confidence.

**Detection Confidence**

As predicted, the relationship between detection confidence and truth bias was positive and significant (p=0.001). Consistent with prior research, the level of confidence subjects had in their assessments of truthfulness positively affected their reliance on the truth bias as a simplifying heuristic (McCornack and Parks, 1986). Subjects who were more certain of what they believed their partners were going to do devoted less cognitive resources toward making veracity judgments. In this study, more detection confidence was shown to come from levels of perceived relational closeness (H3), knowledge of a person’s prior baseline behavior (H2), being proximally located (H6), and the type of communication media used (H5). Consistent with cognitive bias research, subjects inherently proscribed to the “cognitive miser” paradigm and attempted to reduce mental effort by using simplifying heuristics when possible. Subjects felt comfortable relying on the truth bias as a way of making veracity judgments about their communication partners because of their confidence in their decision.

**Truth Bias**

Consistent with prior research, truth bias had a negative relationship with deception detection and false accusations. Using the assumption of truthfulness, individuals systematically mislabeled deceptive statements as truthful statements. Conversely, truth bias also had a positive relationship with truth detection and successful deception. Individuals who relied more on the truth bias to make veracity judgments were more likely to identify deceptive statements as truthful. This study adds to existing deception by showing that the truth bias is robust in computer-mediated environments, across varying types of studies, and differing operationalizations of the construct.
In fact, this study matched all nine of Miller and Stiff’s (1993) criteria for an ideal deception detection study. This study also showed that prisoner dilemma scenarios are an acceptable method of studying deception detection and the truth bias without the negative side effects mentioned by Miller and Stiff. It is also interesting to note that there were almost twice as many truth bias errors (N=21) as there were lie bias errors (N=13).

**LIMITATIONS AND IMPLICATIONS**

This study is limited in that subjects were offered monetary rewards from $0 to $8. Subjects indicated that they viewed these rewards as nominal and, consequently, may not have been motivated enough to exert significant amounts of motivational effort to affect changes in cognitive bias. However, even with small financial incentives, subjects did attempt to deceive their assigned partner and detect deception from that partner. Increased monetary rewards should lead to more motivational effort and more deceptive behavior. In fact, in transcribed recordings and chat logs subjects indicated that they would willingly deceive even close friends for $25.

This study may also have limited generalizability due to differences between student subjects and workers. However, given the general nature of the deception detection process it is unlikely that there will be significant variation between students and workers.

**Implications for Researchers**

This study has three main implications for researchers. First, this study furthers our understanding of new communication media because it shows that richer media may provide more confidence in predicting behavior. However, richer media do not necessarily improve predictive accuracy. In the context of deception, researchers must account for the fact that computer-mediated communication may lead to comparable deception detection rates, compared to face-to-face communication, without additional detection confidence.

Second, researchers gain an understanding of how distributed communication can lead to low levels of detection confidence. The widespread use of new communication media has allowed real-time communication across vast distances, yet the impact of proximity on deception detection has not been investigated. Researchers must recognize that the presence of another individual is a vital factor in studying deception.

Third, researchers must recognize that familiarity and intimacy are separate constructs that both contribute to detection confidence. Researchers also need to be aware that individuals who have known each other for even a relatively short period of time are less accurate at detecting deception than complete strangers. Future studies using subjects that know each other must account for the familiarity and intimacy between subjects.

**Implications for Practitioners**

This study has three main implications for practitioners. First, practitioners need to understand that richer communication, such as face-to-face meetings, can actually lead to less deception detection compared to traditional face-to-face interactions. While richer communications may not directly affect an individual’s ability to detect deception, the findings indicate that they may reduce the individual’s ability indirectly by mistakenly relying on the truth bias. Conversely, using less rich media may comparatively increase deception detection compared to face-to-face communication. This means that corporations...
Distributed Deception

could use non-traditional communication media such as e-mail, voice mail, and virtual meetings to boost performance and, actually, be less susceptible to deception. Improved deception detection within corporations would be valuable for both employers and stakeholders.

Second, individuals can expect higher detection confidence and worse veracity judgments when proximally located to the person with whom they are communicating, regardless of the communication medium. Managers can be confident that increasing utilization of distributed teams for productivity enhancements will not increase susceptibility to deception. In fact, it will lead to relatively higher deception detection rates compared to proximally located teams. However, it may also lead to more false accusations.

Third, practitioners need to understand the impact of working relationships and relational closeness on deception detection. While it may not be possible to eliminate familiarity (baseline knowledge) from the workplace, it would be beneficial if businesses limited the amount of relational closeness among co-workers. By limiting relational closeness and, to the extent possible, familiarity, practitioners will indirectly increase their ability to detect deception.

Future Research

Results from this study have generated several new research questions and directions for future research. First, further research exploring the impact of suspicion, trust, and deception detection in computer-mediated environments, is clearly warranted. The inclusion of these closely-related constructs in a larger model is likely to lead to a clearer picture of how they affect deception detection rates. Examinations of workplace deception within a small group context would also yield valuable clues relative to the impact of deception detection beyond dyadic communication.

Second, future research would benefit from iterated prisoner dilemma sessions with different partners at both the group and individual level. Iterated prisoner dilemma games played in both computer and manual conditions would yield valuable insights into the effect of media richness on negotiation, deception, and strategic decision making. Future research could also include a comparison of different monetary rewards on deception attempts and an individual’s ability to detect deception.

Third, future research should focus on measuring the influence of both situational veracity judgments and past experience. Past research has been driven by the assumption that there is a set of identifiable cues that are specifically related to deceptive behavior (Feeley and Young, 1998). However, the relationship between deceptive cues and pre-established valuations of personal truthfulness has not been investigated. Current knowledge would benefit from additional research aimed at determining the relative predictive powers and interaction of these two factors.

CONCLUSION

This study has yielded several valuable insights into the effects of computer mediation, distributed communication, and knowledge of baseline behavior on deception detection and the truth bias. This study showed that more detection confidence can come from knowledge of a person’s baseline behavior, being proximally located, the type of communication media used, and perceived relational closeness. These factors lead to less deception detection through more detection confidence and reliance on the truth bias. Knowledge of an individual’s baseline behavior also had a direct negative relationship with deception detection. This study found that subjects with a high level of confidence in their ability to detect truthful/deceptive behavior were more reliant on the truth bias to make veracity judgments even in a computer-mediated environment.
REFERENCES


Storms, M.D. (1973). Videotape and the attribution process: Reversing actors’ and observers’
Distributed Deception


APPENDIX

Relational Closeness Scale (RC)

1. X does not influence everyday things in my life. *†
2. X affects my romantic relationships.*
3. X influences my plans to join a club, social organization, church, etc.*
4. X affects my school related plans.
5. X influences my plans for achieving a particular standard of living.*
6. X does not influence how I choose to spend my money. †
7. X influences the way I feel about myself.
8. X does not influence the opinions that I have of other important people in my life. †
9. X does not influence when I see, and the amount of time I spend with, my family. †
10. X influences when I see, and the amount of time I spend with, my friends.*
11. X does not influence which of my friends I see. †
12. X influences how I spend my free time.*
13. X influences when I see X and the amount of time the two of us spend together.
14. X influences what I watch on TV.

* Included in final model.
† Reverse coded.

Communication Media (CM)

1. When we disagreed, the communication conditions made it more difficult for us to come to an agreement.
2. The conditions under which we communicated helped us to better understand each other.* †
3. The conditions under which we communicated slowed down our communications.*
4. When we disagreed, our communication environment helped us come to a common position. †
5. I could easily explain things in this environment.* †
6. The communication conditions helped us exchange communications quickly.* †
7. There were ideas I couldn’t relate to the other party because of the communication conditions.*
8. The conditions under which we communicated got in the way of our sharing of opinions.*

* Included in final model.
† Reverse coded.

Truth Bias (TB)

1. I believe what my partner says with little doubt.
2. I think my partner is generally honest.
3. Overall, my partner was truthful.
4. Overall, my partner was very deceptive. †

† Reverse coded.
Chapter 5.21
Extrinsic Plus Intrinsic
Human Factors Influencing the
Web Usage

Manuel Jesús Sánchez-Franco
University of Seville, Spain

ABSTRACT
Researchers are confronted with a choice among models to explain Web acceptance and usage. Therefore, as Venkatesh et al. (2003) suggest, there is a need for a review and synthesis in order to progress toward a unified view of Web acceptance and usage. In this situation of development, a theoretical model based on technology acceptance (TAM) and flow, is proposed to describe the extrinsic and intrinsic motivations, respectively, for Web users. The aim of the research is thus (1) to investigate how well flow-model theory can be aligned with TAM and (2) to provide a relationship with the Web acceptance and its proper usage. Furthermore, better measures for predicting and explaining Web use will have greater practical value. Singular Web sites would like to assess user demand for new design ideas to facilitate electronic service quality and flow. Users would like to find a Web site leading to an enduring and cost-effective relationship.

INTRODUCTION
The Web can be conceived as a media for contents to build and maintain individualised relationships with profitable customers through its proper usage. In this e-CRM (electronic Customer Relationship Management) context, the viewpoint must change from (1) a traditional perspective with a short-term focus to (2) a long-term perspective with a (2.1) user-retention and (2.2) enduring-involvement focus based on optimal experiences and greater expected Web use to avoid switching suppliers at virtually comparably low direct and indirect costs. Marketers must be interested in users’ profitable sessions at their Web sites being longer and more frequent to increase the degree to which a customer (i.e., Web site user) voluntarily interacts with them. Therefore, it becomes important to examine the human factors that (1) reduce time pressure as a cost for users and (2) influence the acceptance and, in turn, length and
Extrinsic Plus Intrinsic Human Factors Influencing the Web Usage

frequency of Web site sessions (Sánchez-Franco and Rodríguez-Bobada, 2004).

With the growing reliance on computerised systems and increasing speed of the introduction of new Information Technologies (ITs) (e.g., Web), understanding the factors that promote acceptance and effective utilization of Web technology continues to be a vital issue for researchers and practitioners. Specifically, research in HCI (Human-Computer Interaction) tradition has long asserted that the research of human factors (1) is a key to the successful design and implementation of technological devices, and (2) should analyze extrinsic and intrinsic motives. Thus, there is a need for a review and synthesis in order to progress toward a unified view of Web acceptance and usage (Venkatesh et al., 2003).

Accordingly, it is important to consider the human beliefs and affects based on a Technology Acceptance Model (TAM) (i.e., ease of use and usefulness) and a flow model, respectively, to understand: (1) attitude towards using the Web; (2) behavioural intention to use; and (3) Web actual usage. On the one hand, the two beliefs based on TAM, perceived usefulness and ease of use, are the most important human factors determining usage of computers or IS (Information System). On the other hand, flow, defined as an optimal, intense and intrinsically enjoyable experience, has been proposed as a useful framework (1) for studying the experience of individuals as they learn and use the Web, (2) for identifying the factors that influence this experience, and (3) as a way of defining the nature of compelling online experiences (Novak et al., 2000). In fact, creating compelling experiences in this distinctive consumption environment depends on facilitating a state of flow (Csikszentmihalyi, 1975, 1990; Hoffman and Novak, 1996b; Novak et al., 2000).

However, very little is theoretically and empirically known about users’ interactions with Web-based technologies. Few studies actually focus directly on (1) Web acceptance and usage and its antecedents and consequences adopting a user-centered perspective, and (2) the extrinsic and intrinsic motives that affect Web usage. In this sense, Novak et al. (2000) suggest that among marketing academics and Internet practitioners alike, there is a lack of genuine knowledge about the factors that (1) make for effective interactions with online users and (2) make using the Web a compelling user experience. More recently, Parasuraman and Zinkhan (2002) point out that there is a considerable knowledge gap between the practice of online marketing and the availability of sound, research-based insights and principles for guiding that practice.

In this situation of development, a theoretical model based on technology acceptance (TAM) and a flow model (concerning an Information System), is proposed in this chapter to describe the extrinsic and intrinsic motives for online users. Chan and Swatman (2000) stated that there is very little literature which discusses the process of Internet-based marketing, so that researchers must (1) start with the literature concerning more general IS implementation and (2) hope to develop a body of theory, which is more explicitly focused on the area of Internet marketing (Eid and Trueman, 2002). Our objective in this chapter is thus to evaluate the mediating role of main intrinsic and extrinsic motives explaining users’ Web acceptance and affecting the Web usage (1) to explain and (2) to improve the users’ experience of being and reacting in the Web, and, in turn, (3) to run a profitable business.

THEORETICAL FOUNDATIONS

Over the last two decades, a significant body of research has focused on identifying various factors that influence user-acceptance behaviour, putting forward several theoretical models. In particular, the Technology Acceptance Model (TAM), introduced by Davis and his colleagues (Davis, 1989; Davis et al., 1989), has received considerable attention (see Lucas and Spitler, 1999, for a
review). Several researches have demonstrated the validity of this model across a wide variety of IS (see Moon and Kim, 2001). Specifically, the model was shown to have good predictive validity for the use of several ITs including e-mail and the Web (Fenech, 1998; Gefen and Straub, 1997).

It has thus become established as a parsimonious yet powerful model for (1) explaining attitude towards using IS, and (2) predicting usage intentions and its adoption. In other words, to understand (1) the causal link between external variable and user acceptance of PC-based applications (Fenech, 1998); and, more recently, (2) human Web acceptance (Johnson and Hignite, 2000; Lin and Lu, 2000).

**Technology Acceptance Model**

Davis adapted the Theory of Reasoned Action (TRA) to TAM by developing two key beliefs (i.e., usefulness and ease of use) that specifically account for IS usage as a basic dependent variable of IS. TAM adopts the well-established causal chain of beliefs → attitudes → intention → behaviour (TRA) that has been put forward by social psychologists, Fishbein and Ajzen (Ajzen, 1991; Fishbein and Ajzen, 1975). Consistent with TRA, both users’ beliefs determine the attitudes toward using the system. Behavioural intentions to use, in turn, are determined by these attitudes toward using the system. Finally, behavioural intentions to use lead to actual system use (see Figure 1).

The first of these main beliefs is perceived usefulness. It is defined as “the degree to which a person believes that using a particular system would enhance his or her performance” (Davis, 1989). Perceived Usefulness was originally seen as a fairly simple concept including components of effectiveness and efficiency (that are mainly related to extrinsic motivation in work contexts). As shown by Davis (1989), perceived usefulness affects usage of computers. Specifically, Teo et al. (1999) found that perceived usefulness has a strong significant relationship with the Web usage. For example, e-shoppers will use the Web

**Figure 1. TAM**

(*) Attitude was originally included as a mediator between the personal belief-based constructs and the behavioural intention (Davis et al., 1989), but later was dropped from the model because it was finally found to be a weak mediator (Davis et al., 1992).

(**) Some researchers (see, for example, Thompson et al., 1991; Adams et al., 1992; Al-Gahtani and King, 1999) also dropped the behavioural intention to use because they were interested in actual behaviour (system usage) and not intentions.
sites more if they find them useful for shopping offering quality information helpful for shopping as well as useful functionality (such as online order status tracking capability, Baty and Lee, 1995; Bellman et al., 1999). In short, individuals will use IS if they perceive that such usage would help them to achieve and enhance the desired task performance, even if it is at first difficult to use (Eid and Trueman, 2002).

The second belief is perceived ease of use, defined as “the degree to which a person believes that using a particular system would be free of effort” (Davis, 1989), being determined by the users’ skills and the usability of the system (Venkatesh and Davis, 1996). Perceived ease of use has been (1) used as a measure of system quality in studies of IS success (Seddon, 1997); (2) considered a component of a Web site’s system quality (Liu and Arnett, 2000); and (3) found to influence computer usage and the Internet usage indirectly via (3.1) perceived usefulness (Davis, 1989; Teo et al., 1999) and (3.2) perceived enjoyment (Igbaria et al., 1995; Teo et al., 1999). Thus, ease of use is an important component when measuring user satisfaction with a Web site (Wang et al., 2001) and its usage (Elliot and Fowell, 2000).

In this context, as perceived ease of use has an inverse relationship with the perceived complexity of use of the technology, it can affect perceived usefulness. A system that is difficult to use is less likely to be perceived as useful. Assuming other things being equal, users consider a system more useful when it is more effort-free. These relationships have been examined and supported by many prior studies (Davis, 1989; Davis et al., 1989; Venkatesh and Davis, 1996; Venkatesh and Davis, 2000). Moreover, if the challenges of an activity are beyond the individual’s skill level, demanding more than the individual can handle, a state of anxiety ensues and users (1) interpret challenges as simply functional complexity or obscurity (and not as opportunities for action), (2) do not perceive the system as useful, and thus (3) tend to use the system sporadically. Likewise, a system perceived as difficult to use is less likely to be perceived as enjoyable, leading to decreased usage (Lim, 2002). Thus, from a causal perspective, ease of use may be an antecedent to usefulness rather than a parallel determinant of usage.

In short, as Hubona and Geitz (1997) reported, perceived usefulness and perceived ease of use have sound theoretical foundations. They are therefore widely accepted as valid and predictive measures of future Web usage levels. TAM yields highly consistent results in the acceptance behaviour of the users towards new systems.

However, most of the TAM research has only been conducted from an extrinsic motivation perspective (Igbaria et al., 1996). Researchers have become increasingly aware of the relevance of the non-cognitive aspects of use motives such as emotions, symbolism and hedonistic desires in understanding facets of behaviour. For example, user behaviour-based findings in the intrinsic motivation and self-efficacy research indicate that emotional responses play important roles in determining (1) a person’s attitude towards using the Web, (2) a behavioural intention, and (3) an actual behaviour.

Following Human-Computer Interaction (HCI) Research, several researchers propose the need for incorporating intrinsic human factors or integrating other theories in a specific study to improve its particular and explanatory value (Hu et al., 1999; LeGris et al., 2003; Venkatesh and Davis, 2000). For example, psychologists have proposed a variety of theories explaining how behavioural reactions are influenced both by cognition and affect (Berkowitz, 1993; Epstein, 1994; Leventhal, 1984; Zajonc, 1980). Specifically, one of the intrinsic human motives related to prior factors is “flow.” Next we analyse the proposed flow models.

Flow Theory

Flow has been particularly studied in the context of ITs and hypermedia computer-mediated
### Table 1. Summary of research on flow

<table>
<thead>
<tr>
<th>Authors</th>
<th>Dimensions</th>
<th>Antecedents</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghani, Supnick and Rooney (1991)</td>
<td>Concentration</td>
<td>Individual Skills</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Enjoyment</td>
<td>Control</td>
<td>Challenge</td>
</tr>
</tbody>
</table>

Ghani et al. (1991) argued that two key characteristics of flow are (1) total concentration in an activity and (2) the enjoyment one derives from an activity. Control and flow predicted exploratory use, which in turn predicted extent of use.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Dimensions</th>
<th>Antecedents</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trevino and Webster (1992)</td>
<td>Control</td>
<td>Computer Skill</td>
<td>Attitudes</td>
</tr>
<tr>
<td></td>
<td>Attention Focus</td>
<td>Technology Type</td>
<td>Effectiveness</td>
</tr>
<tr>
<td></td>
<td>Curiosity</td>
<td>Ease of Use</td>
<td>Quantity</td>
</tr>
<tr>
<td></td>
<td>Intrinsic Interest</td>
<td></td>
<td>Barrier Reduction</td>
</tr>
</tbody>
</table>

Trevino and Webster (1992) described four dimensions of the flow experience in the context of ITs: (a) the user perceives a sense of control over the computer interaction, (b) the user perceives that his or her attention is focused on the interaction, (c) the user’s curiosity is aroused during the interaction, and (d) the user finds the interaction intrinsically interesting, implying that the user’s interaction with the technology extends beyond mere instrumentality, becoming a pleasure and enjoyable as an end in itself.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Dimensions</th>
<th>Antecedents</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Webster, Trevino and Ryan (1993)</td>
<td>Control</td>
<td>Perceived Flexibility</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attention Focus</td>
<td>Perceived Modifiability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cognitive Focus</td>
<td>Experimentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Enjoyment</td>
<td>Future voluntary Use</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Actual use</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Perceived Communication Quantity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Perceived Communication Affectiveness</td>
<td></td>
</tr>
</tbody>
</table>

Webster et al. (1993) refined the model to just three dimensions: (1) control; (2) focus attention; and (3) curiosity and intrinsic interest coalescing to become cognitive enjoyment (a construct comprised of curiosity and intrinsic interest that were highly interdependent). Flow would be associated with specific characteristics of the software (specifically, perceptions of flexibility and modifiability) and with certain technology use behaviours (experimentation and future voluntary computer interactions) (Agarwal and Karahanna 2000).

<table>
<thead>
<tr>
<th>Authors</th>
<th>Dimensions</th>
<th>Antecedents</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghani and Deshpande (1994)</td>
<td>Concentration</td>
<td>Skill</td>
<td>Exploratory Use</td>
</tr>
<tr>
<td></td>
<td>Enjoyment</td>
<td>Control</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Challenge</td>
<td></td>
</tr>
</tbody>
</table>

In a later study exploring flow occurring among individuals using computers in the workplace, Ghani and Deshpande (1994) analyzed skill as well as challenge. Skill leads to control which leads to flow. Skill also directly affects flow, as does perceived challenge, with an optimum level of challenge relative to a certain skill level existing. A second factor affecting the experience of flow is a sense of control over one’s environment.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Dimensions</th>
<th>Antecedents</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoffman and Novak (1996b)</td>
<td>Not specified</td>
<td>Primary antecedents:</td>
<td>Learning</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Skills / Challenges</td>
<td>Perceived behavioural</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Focused Attention</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary antecedents:</td>
<td>Positive subjective</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Telepresence</td>
<td>Experience</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interactivity</td>
<td>Distortion in Time Perception</td>
</tr>
</tbody>
</table>

*Adapted from Agarwal and Karahanna (2000) and Sánchez-Franco (2003)*
A significantly more complex version of flow was described by Hoffman and Novak (1996b). Examining the role of marketing in CME, Hoffman and Novak argued that the dimensions of control, curiosity, intrinsic interest, and attention focus were antecedents to flow. Their model included several other antecedents of flow such as the perceived congruence of skills and challenges and the telepresence of the medium, defined as the mediated perception of an environment (Steuer, 1992). Hoffman and Novak indicated that the primary antecedents to flow are challenges, skills, and focused attention.

From the literature on communication media, they added secondary antecedents: (1) interactivity, and (2) telepresence. Furthermore, Hoffman and Novak added the construct of involvement, which encompasses intrinsic motivation and self-reliance and is influenced by whether the activity is goal-directed or experiential (Finneran and Zhang 2002). They further theorised that flow would result in several outcomes such as a positive subjective experience, increased learning, and perceived behavioural control.


<table>
<thead>
<tr>
<th>Authors</th>
<th>Dimensions</th>
<th>Antecedents</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not specified</td>
<td>Primary antecedents: Control Arousal Focused attention Secondary antecedents: Challenge Skill Interaction speed Involvement</td>
<td>Positive affects Exploratory behaviour</td>
<td></td>
</tr>
</tbody>
</table>

More recently, Novak et al. (2000) took the definition of flow to the operational level (in a CME), stating that flow is “determined by: a) high levels of skill and control; b) high levels of challenge and arousal; c) focused attention; and (…) d) enhanced by interactivity and telepresence.” Thus, flow occurs when an activity challenges and interests individuals enough to encourage (1) playful and exploratory behaviour without the activity being beyond the individuals’ reach, and (2) greater expected Web use.

*Adapted from Agarwal and Karahanna (2000) and Sánchez-Franco (2003)*

Flow, defined as “the holistic sensation that people feel when they act with total involvement” (Csikszentmihalyi, 1975), has been recommended as a possible metric of the online consumer experience (Agarwal and Karahanna, 2000; Ghani et al., 1991; Ghani and Deshpande, 1994; Hoffman and Novak, 1996b; Novak et al., 2000; Trevino and Webster, 1992; Webster et al., 1993). Therefore, we suggest that flow-based theory could contribute partly to explain attitudes towards using the Web-based technologies and behaviours.

Although a body of research suggests that flow on the Web is fleeting, rarely experienced, associated with the increases in depression and loneliness (see Kraut et al., 1998), and mostly by novice Web users, the growing research concerning theory of optimal flow has been proposed (1) as a useful framework for identifying the factors that influence this experience and, in turn, (2) as environments (CMEs, defined by Hoffman and Novak, 1996b, as a distributed computer network used to access and provide hypermedia content).
Extrinsic Plus Intrinsic Human Factors Influencing the Web Usage

a way of defining the nature of compelling online experiences (Table 1).

Next we analyse the main factors that influence this optimal experience. We distinguish two submodels for better understanding of the relationships between TAM-based beliefs and flow state.

Submodel 1

Researchers have maintained that involvement is a major socio-psychological variable that explains individual differences (Petty et al., 1981) that impact on attitude and individual behaviour. Following a review of the construct of involvement in psychology, organisational behaviour, and marketing, Barki and Hartwick (1989) conclude that these disciplines have converged in a definition of involvement “as a subjective psychological state, reflecting the importance and personal relevance of an object or event.” In this cognitive-processing context, involvement has also been argued to have a significant effect on consumer subjective perception of how much they think they know about products (Zinkhan and Muderrisoglu, 1985). In turn, Houston and Rothschild (1978) and others have found that involvement increases with familiarity with the stimulus or individual’s prior knowledge (i.e., ability) (Figure 2.)

On the other hand, involvement contributes to the attention focused on the stimulus (Zaichkowsky, 1986) and it is considered as a prerequisite for flow (Hoffman and Novak, 1996b). Ghani and Deshpande (1994) emphasise that the total concentration in an activity is the key characteristic of flow (Figure 2). According to Csikszentmihalyi and Csikszentmihalyi (1988), when one is in flow “one simply does not have enough attention left to think about anything else.” The computer screen functions as a limited stimulus field. Moreover, involved users report being mesmerised during their computer interactions. Accordingly, Park and Young (1986) note that users — for whom extrinsic motives are salient — focus their attention on utilitarian cues and evoke cognitive responses. In turn, users — for whom intrinsic motives are salient — focus their attention on symbolic or experiential cues and evoke emotional responses.

Involvement can be thus understood by distinguishing the types of involvement according to the motives underlying involvement (Park and Young, 1986). Particularly, the distinction between extrinsic and intrinsic motives of behaviour suggests two types of involvement:

- Situational involvement arises from several transitory factors that affect the relationship between the individual and the stimulus (Celsi and Olson, 1988). It is externally motivated and it is thus more likely to result in a goal-directed behaviour (Hoffman and Novak, 1996b). Extrinsic motives-based involvement (i.e., SI) is related to the performance of an activity. In this sense, the activity is perceived to be instrumental in achieving valued outcomes that are distinct from the activity itself.

- Enduring involvement is an intrinsically-motivated individual difference variable that is relatively long-lasting. In this context, intrinsic motivation refers to the performance of an activity for no apparent reinforcement other than the process of performing the activity per se. According to no apparent reinforcement, the concept of perceived enjoyment is defined as “the extent to which the activity of using the computer is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated” (Davis et al., 1992). In short, intrinsic motives to use the Web could be associated with frequent Web use for intrinsically enjoyable purposes.

Therefore, situational involvement reflects temporary feelings of involvement that accompany a particular situation, whereas enduring involvement is an individual difference variable
representing the general, long-run concern with a stimulus that a consumer brings to a situation (Richins et al., 1992).

According to intrinsic motives related to enduring involvement, enjoyment has been identified as an important motivational factor in computer use, (1) contributing towards creativity and exploratory use behaviour (Ghani, 1991), as well as (2) being a major dimension of optimal experience or flow, which has been above described as an intrinsically enjoyable experience (Csikszentmihalyi, 1975). Specifically, research on the use of the Web has found empirical support for enjoyment as a driver of Web usage (e.g., Atkinson and Kydd, 1997; Moon and Kim, 2001; Teo et al., 1999). If individuals like and enjoy their Web browsing experience, it is likely that they are going to (1) involve in browsing and, in turn, (2) enhance their online service perceptions (e.g., perceived usefulness and ease of use). Use of the Web may therefore evoke emotional values that are not only captured by ease of use or usefulness (Hoffman and Novak, 1996ab; Singh and Nikunj, 1999). Use of the Web goes beyond utilitarian aspects to include intrinsic enjoyable experience (Berthon et al., 1996; Pine and Gilmore, 1998). For example, Davis et al. (1992) argued that “while usefulness will once again emerge as a major determinant of intentions to use a computer in the workplace, enjoyment will explain significant variance in usage intentions beyond that accounted for by usefulness alone.”

Moreover, HCI-based research using the TAM model has found that perceived enjoyment of using a system (e.g., Web) has a relationship with a perceived ease of use (Venkatesh, 1999; Venkatesh, 2000; Moon and Kim, 2001) and perceived usefulness of the system (Agarwal and Karahanna, 2000) (Figure 2).

On the one hand, Agarwal and Karahanna (2000) found a multi-dimensional construct called cognitive absorption (similar to flow state) which had a significant influence on usefulness over and above ease of use. Venkatesh (2000) showed that enjoyment influenced usefulness via ease of use without assessing its direct effect on usefulness over and above ease of use (Yi, 2003). Likewise, several researchers note that when the usage experience is more enjoyable the impact of perceived usefulness on Web usage could be relatively lower. This prior phenomenon is based on a cognitive consistency argument in which the underlying rule is that when usage is emotional, instrumental issues -such as perceived usefulness- ought not to come into one’s main decision making criteria for future usage (Chin et al., 1996). However, the

---

Figure 2. Extending model (I)

*Flow has been defined as intrinsically enjoyable experience.*
Extrinsic Plus Intrinsic Human Factors Influencing the Web Usage

effect of enjoyment on perceived usefulness is still relatively unknown.

On the other hand, Csikszentmihalyi (1975) argued that flow could be enhanced when an individual perceived an activity to be executed easily. Empirical research has also found support for this relationship in traditional settings (Igbaria et al., 1996). It is conceivable that a Web site that is easier to use provides better feedback to a visitor's stimuli, and consequently, leads to increased enjoyment and flow. Moreover, Venkatesh (1999) compared two training methods (traditional training vs. game-based training) and found that the training method with a component aimed at enhancing intrinsic motivation induced higher ease of use perceptions. Later, as we commented above, Venkatesh (2000) conceptualised enjoyment as an antecedent of ease of use, whose effect increases over time as users gain more experience and perceived control with the system (adapted from Hwang and Yi, 2002).

Finally, studies applying the perspective of flow have shown that to provide intrinsic motivation, some services must represent a certain challenge to the user as antecedent of emotional arousal. It is probably that excessive ease of use that reduces the sense of accomplishment, (1) negatively influences on perceived enjoyment and (2) leads to boredom states. In other words, ease of use is not the only key criterion for Web site design, as Web site usage would decrease with time. On the contrary, a main determinant must be its stimulating use, so that it evokes compelling experiences and therefore increases profitable Web site use (Sánchez-Franco and Rodríguez-Bobada, 2004). According to Csikszentmihalyi (1996), a Web site must be challenging, competitive, and provide feedback to its users in order to encourage the occurrence of flow. As Ginzberg (1978) recommended, system success must be evaluated in terms of the way it is used rather than just the extent of use. Therefore, an important prerequisite for this rewarding experience is that an individual is able to accomplish the task (i.e., ability). But it is equally important for it to be experienced as a challenge and the individual gets (1) stimulation (i.e., arousal) and (2) unambiguous feedback (i.e., perceived control) inherent in the performance of the activity. To complete the global model, we thus introduce a second submodel based on users’ ability, challenge, arousal and perceived control.

Submodel 2

Looking at earlier definitions of flow (Table 1), in order to experience flow while engaged in an activity, users must perceive a balance between their abilities (defined as the skill or proficiency) and the challenges of the activity (defined as their opportunities for action on the Internet) (Novak et al., 2000). Particularly, challenges are related to a sense of accomplishment rather than simply functional complexity or obscurity. Both their abilities and challenges must be above a critical threshold (Massimini and Carli, 1988). The balance facilitates the experience of arousal, perceived control and flow.

- If the challenges of an activity (i.e., those opportunities which provoke users to further explore Web sites) demand more than the individual can transitorily handle, a state of stimulation ensues (i.e., arousal: high challenge/ moderated skill). Users become aroused until they are familiarised with the system through more frequent system usage –e.g., practice or training (Gardner et al., 1989) (Figure 3). However, too much stimulation will lead users to making errors and feel out of control (i.e., anxiety as a negative affective reaction toward Web use).
- On the contrary, when the challenges are lower than the individual’s skill level, perceived control (moderated challenge/high skill) may be the result (Figure 3). However, if the challenges are too low, users lose in-
Interest and tend to use the Web sporadically (i.e., boredom or apathy).

According to arousal as an emotional response, it reflects a user's concern about having the ability to succeed with a new perceived challenge. Arousal can be thus considered as an involvement-based response. However, the user's concern must be perceived as moderate, important and relevant (1) to acquire Web-based skills and (2) to match the skill level and perceived challenge. On the contrary, too much concern will lead users to feel out of control.

Moreover, if users return to the same Web site over time, it is reasonable to expect (1) learning to occur, (2) perceived challenges to decrease, and (3) session lengths to decline (see, for example, Johnson et al., 2003, on “the power law of practice”). A main recommendation must be thus to promote its stimulating use, so that it permanently evokes (1) arousal and compelling experiences, and (2) more frequent and longer visits. For example, CMC (Computer-Mediated Communication) technologies can stimulate cognitive curiosity and the desire to attain competence with the technology by providing options such as menus that encourage exploration (Malone and Lepper, 1987) and competence attainment. In this sense, arousal, as a consequence of perceived task challenge, is a key factor in the experience of flow.

According to perceived control, it has been studied in the context of electronic commerce and found to have a positive effect on customer attitudes and behaviour (Ghani et al., 1991; Novak et al., 2000; Koufaris et al., 2001-2002). Specifically, it refers to users' perception of their capabilities to interact in CME. Perceived control comes from: (1) the users’ perception of their ability to adjust the CME; (2) their perception of how the CME responds to their input; plus (3) an environment where challenges are relatively moderate. These users thus believe their actions and abilities determine their successes or failures (Sánchez-Franco and Rodríguez-Bobada, 2004).

Therefore, users with a high level of ability and, consequently, perceived control (Ghani and Despandhe, 1994; Novak et al., 2000): (1) are likely to feel more able to perform the activity,
Extrinsic Plus Intrinsic Human Factors Influencing the Web Usage

and (2) show a high comfort level. They would be more inclined to feelings of enjoyment while become involved in the activity and, in turn, to use the Web more frequently. Likewise, users become more playful (Lieberman, 1977) and experiential, positively affecting Web exposure length. As Bandura (1982) suggested, “people do not perform maximally, though they possess the constituent skills.” He suggests that “the reason people enjoy challenging tasks is that by testing the upper limits of their competencies, they find out what they are able to do, thereby increasing their feelings of self-efficacy.” On the contrary, those with low self-efficacy expectations in a particular situation will experience unpleasant feelings, such as anxiety, and will behave in unproductive ways, such as avoiding work, and may lack persistence (Bandura, 1977).

Perceived control is thus similar to Bandura’s self-efficacy (1982) defined as “judgements of how well one can execute courses of action required to deal with prospective situations.” It is (1) specific to an action and it can be different across situations or actions; (2) facilitated by the medium adapting to feedback from the individual, and by providing explicit choices among alternatives (Webster et al., 1993), and (3) considered -by several researchers- as an antecedent of perceived behavioural control (i.e., “perceptions of internal and external constraints on behaviour,” Taylor and Todd, 1995).

Finally, perceived control (or self-efficacy) can be related to perceived ease of use. Users regard the system easier to use when their conviction in their own efficacy regarding the particular system is higher (Agarwal et al., 2000; Venkatesh and Davis, 1996; Venkatesh, 2000).

Implications for Web Usage

Based on the above theoretical development (Figure 3), two main Web user-types can be theoretically evidenced on a continuum from “pure browsing” to “pure seeking.” The distinction is a continuum rather than a dichotomy. Individual differences drive a person’s information and entertainment consumption processes. Individuals shift from one mode to the other.

When users show ritualised orientations exploring the Web (experiential behaviour), they are moved by an intrinsic motive: “to feel pleasure and enjoyment from the activity itself” (Bloch et al., 1986). Users find the interaction intrinsically interesting. They are involved in the activity for the emotional responses it provides rather than for utilitarian purposes. Thus, a main objective is that a Web site is designed to be stimulating to use and thus to evoke compelling user experiences related to playfulness, exploratory behaviour and positive affects.

According to playfulness and positive affects, Atkinson and Kydd (1997) examined the influence of playfulness on the use of the Web, defined as the degree of cognitive spontaneity in microcomputer interactions (Webster and Martocchio, 1992). They found that both playfulness and usefulness affect its use in different ways, depending on its use for entertaining or for work. Likewise, they found that playfulness is significantly associated with total Web use. Those who are more playful with computers tend to indulge in using a new system just for the sake of using it. Therefore, they in general underestimate the difficulty associated with using a new system (Venkatesh, 2000). Previous computer adoption studies have verified that if users are more playful with computer systems, they are more willing to use the systems (Igbaria et al., 1994; Webster and Martocchio, 1992). In turn, Webster et al. (1993) note that research has suggested that “higher playfulness results in immediate subjective experiences such as positive mood and satisfaction” (Levy, 1983; McGrath and Kelly, 1986; Sandelands et al., 1983). Furthermore, previous research on human-computer interactions (Sandelands and Buckner, 1989; Starbuck and Webster, 1991; Webster and Martocchio, 1992) has shown that higher degrees of pleasure and involvement during computer interactions lead
to concurrent subjective perceptions of positive affect and mood (Hoffman and Novak, 1996b).

Also, Amabile (1988) noted that “only the intrinsically motivated person (…) who is motivated by the interest, challenge, and the enjoyment of being in the maze (…) will explore, and take the risk of running into a dead-end here and there.” In this sense, Ghani and Deshpande (1994) examined flow in the context of individuals who used computers in their daily work and found that it had a significant impact on exploratory use of the computer which, in turn, had a significant effect on the extent of computer use.

When users show an instrumental orientation to the Web (goal-directed behaviour), they search for contents adapted to their needs and goals and leave the Web after an active and efficient search. Pure seekers use the Web less for experiential activities and more for goal-directed activities based on perceived usefulness. Thus, an objective is that a Web site is designed to be easy to use and useful to increase profitable Web site usage. Likewise, goal-directed users are generally involved in activities that already have a high extrinsic motivating potential. Such individuals are less likely to seek challenges and evoke arousal in Web use (Ghani and Deshpande, 1994). As users become more skilful, their information search shifts from an extensive manner to a simplified one. Web users can evidence opportunity costs of time and confront a variety of time constraints. An idea put forward in some early empirical research on the Web holds that Web users will continue to browse as long as the expected benefit or value of an additional page view exceeds the cost (Sánchez-Franco and Rodríguez-Bobada, 2004).

In short, as Novak et al. (2000) suggest, because the Web mixes experiential and goal-directed behaviours, the model can be used as a first step in evaluating Web sites in terms of the extent they deliver these two types of experience while users browse.

**FUTURE TRENDS**

The objective of this chapter was to create an opportunity for researchers desiring to advance Web-adapted theory. However, due to the limitations of this chapter, there are still some related problems that should be investigated in the future. The specific variables in each category of our framework are not exhaustive, but reflect factors that the literature suggests are most likely to be relevant to Web acceptance research and implementation. Research questions can be raised according to the major framework components.

We argue that the relationships between affective and cognitive dimensions forming both electronic service quality and satisfaction may provide useful insights for how firms should allocate resources in different psychographic segments depending on browsing behaviour. In future research, we must analyse if the relative importance of the electronic service quality and satisfaction dimensions varies from one segment to another. Therefore, service firms may benefit from allocating their resources differently in each segment.

Likewise, we restricted our investigation to intrinsic and extrinsic motives. However, user behaviour is explained via a model of triadic reciprocity in which behaviour, cognitive and personal factors and environmental events all operate interactively as determinants of each other (Hwang and Yi, 2002). Thus, it is necessary to fulfil the model with the role of consumer demographic variables and navigation context (work/home, high/low download, etc.) that are unexplored in this research. Likewise, to the extent that using a CME depends on non-motivational factors like “requisite opportunities and resources” (for example, Internet access), the traditional formulation will not accurately predict intentions and subsequent Web usage (Hoffman and Novak, 1996b).
Finally, the present research may be effectively extended beyond (1) a general analysis of Web acceptance and usage to the modelling in specific Web sites and (2) products/services suited to a relationship-oriented market approach. The research cannot provide insights into the role of personal factors on individual sites. These factors can play distinct roles depending on the nature of the Web site, or nature of the visit.

CONCLUSION

This chapter examines two theories that have been widely used over the past decade to assist in understanding of the IS/IT adoption and implementation processes, and links them to the marketing value chain. The theoretical development suggests that there is scope for further extension of TAM to adapt to the Web-based usage and its profitable consequences. Therefore, placed in this context, the theory may help to further the empirical research and to clarify and examine a Web acceptance and usage model. In other words, the aim of the research has been to investigate how well flow-model theory can be aligned with TAM and to provide a relationship with the Web acceptance and its proper usage.

In this e-CRM context, the design of a Web site based on technology acceptance and flow theories is a crucial determinant of whether visitors are likely to accept and return to the site. An effective site design has a significant influence on site traffic and the number of transactions users conduct on a Web site. A main conclusion is that a Web site must not be simply designed as easy to use thus decreasing Web site usage (aspects related to SI). It is designed (1) to be stimulating to use and thus (2) to evoke compelling user experiences to increase profitable Web site usage. The online environment must promote a long-term perspective with a user-retention focus and involvement based on extrinsic and intrinsic motives, leading to valuable behaviours and, in turn, greater expected Web use. Therefore, flow would occur during both goal-directed as well as experiential types of activities.

Likewise, rather than just evaluating the theoretical aspects, the basic model can be used to assess motivational design aspects during the browsing process. Exploring flow-antecedents and consequences for predicting and explaining Web-technology acceptance and usage would have greater practical value. (1) Web sites would be likely to assess user demand for new design ideas to facilitate flow. (2) Users would be likely to find a Web site leading to an enduring relationship and lessen users’ boredom and anxiety.

REFERENCES


Extrinsic Plus Intrinsic Human Factors Influencing the Web Usage


Venkatesh, V. (1999). Creation of favorable user perceptions: Exploring the role of intrinsic motiva-
Extrinsic Plus Intrinsic Human Factors Influencing the Web Usage


This section presents contemporary coverage of the managerial implications of human-computer interaction. Particular contributions address virtual networking, knowledge blogs in company environments, and ambient business. The managerial research provided in this section allows executives, practitioners, and researchers to gain a better sense of the relationship between business, technology, and individuals.
Chapter 6.1
Social Impact of Virtual Networking

Hakikur Rahman
SDNP Bangladesh, Bangladesh

INTRODUCTION

Information has been defined as a set of data, facts, and figures that have been processed in such a way that they become meaningful. They make intelligence. When information is applied to doing something and is globally pertinent, it is said to have become knowledge.

Information flow can be treated as an alternate wealth for a developing society and knowledge networking through virtual communication processes can break the lags and leads of information barriers. It can create an appropriate tool for achieving and facilitating exchange of information and knowledge among development partners, academia, policymakers, and the civil society at local, national, and global level to design and implement plans for development (Rahman, 2000).

Virtual communities are the collection of online links to a particular node, examples of which are Yahoo!, eBay, Amazon, or smaller chat rooms or instant message buddy lists. These networks of links are freely chosen, democratic, unrestricted, and may even be anonymous or pseudonymous (Roberts, Smith, & Pollock, 2002).

The concept of the virtual enterprise has emerged in management literature as the result of the fusion of technological advances and a claimed major socioeconomic paradigm shift. The virtual enterprise can be seen as a temporary alliance of contracted individuals or companies linked together by ICTs, which assembles for the purpose of a specific business task. Advocates of the virtual enterprise believe that it will replace the conventional model of organization in the 21st century (Introna, More, & Cushman, 1999).

The virtual network is being increasingly promoted as a model for a new form of ICT-mediated communication endeavor. Initially, the concept of the virtual network and the supportive role of ICTs as conceived by its proponents need to be clarified. Based on the initial understanding, the establishment of community information centres as the existing instance of virtual enterprise needs to be done.
BACKGROUND

A virtual organization is a collection of geographically distributed, functionally, and/or culturally diverse entities that are linked by electronic forms of communication and rely on lateral, dynamic relationships for coordination. Despite its diffused nature, a common identity holds the organization together in the minds of members, customers, or other constituents (DeSanctis & Monge, 1998). Virtual is defined as “being in essence or effect but not in fact or name,” and network as “an interconnected or interrelated chain, group, or system” (Lau & Hayward, 2000, p. 362).

To accelerate the development processes at the marginal communities, network hierarchies must reach remote places with easy access and availability, forming a robust intercommunication network. Figure 1 shows possible networking hierarchy within a country.

Rheingold (1994) defines virtual communities as “social aggregations that emerge from the Net when enough people carry on public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace” (p. 5). Integrating this concept along with economic, politics, and social aspects, networking among the virtual communities can be bonded to form a self-sustained en masse at the outset (Fisher, Sonn, & Bishop, 2002).

Communication is fundamental to any corner of the information coordinates but is preeminent in virtual organizations. In O’Hara-Devereaux and Johansen’s (1994) view, without communication, the boundary-spanning among virtual entities would not be possible. Electronic communication enables parties to link across distance, time, culture, departments, and organizations, thereby creating “anyone/anytime/anyplace” alternatives to the traditional same-time, same-place, functionally centered, in-house forms of organizational experience.

Usually virtual networks are characterized by their (a) highly dynamic nature, (b) vibrant relationships among entities, (c) unrestricted boundaries, and (d) easily configurable structures. Relative to more traditional settings, communication processes that occur in virtual contexts are faster, customized, momentary, greater in volume, more informal, and more relationship based.

Barabasi (2002) presents a set of concepts which, taken together, comprise his science of networks. His theory addresses varied entities, from the microscopic cell to the macroscopic World Wide Web, as networks. It seems possible that his work offers to our discipline a new model, one that may address historic dichotomies of person/community, locale/relationship, gemeinschaft/gesellschaft, and one/many. It may provide a path for virtual communities to become part of our ongoing research and conversation.

Enabling the remotely located dispersed communities with contents of their own need, at the times of their own demand, at the easy reach of their own are a few preconditions to enlighten themselves with knowledge and raise their capacity to contribute for the development
of the society. Therefore, these forms of virtual networks, at the village/community level, create enormous opportunities to enhance the human capacity development processes and at the same time raise the economic platform of the community through diversified activities related to their livelihood.

**METHODOLOGY**

Networking minimizes the task of unoptimized searching and sequencing in distributed environment. Individual stations connected to cluster of networks and eventually interconnected together using similar infrastructures and optimized protocols are becoming popular throughout the virtual communities (Rahman, 2001).

Methods and approaches in organizing networked systems need intensive research, especially when the challenge is to implement information technology methodologies effectively to support organizational systems. Qureshi, Vreede, dG-J., and Vogel (2000) indicate that an understanding of organizational systems may be achieved through the use of research methods that (1) use theories to describe organizational systems, (2) provide sets of tools to enable real-world problems to be addressed, and (3) enable the researcher to interact with the organizational systems that they study.

For advanced information services, the powerful clusters/groups in society are provided with direct access to information. The information is then expected to trickle down to the majority at large. Quite often, research centres, for example,

---

**Figure 2. Top-down information channel**

<table>
<thead>
<tr>
<th>Central node (one at the head quarter, with a backup node)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regional nodes (several dividing the country at a homogenous context, or could be division, province, state)</td>
</tr>
<tr>
<td>City/Town nodes (major cities, strategic towns, district towns, content driven locations)</td>
</tr>
<tr>
<td>Village Information Centres (end user stations used as multipurpose information and service centres; could be in strategic locations in villages, commodity driven locations, agricultural product driven locations)</td>
</tr>
</tbody>
</table>

**Table 1.**

<table>
<thead>
<tr>
<th>Types of end user</th>
<th>Usage pattern and average usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Light use</td>
</tr>
<tr>
<td>Cyber centre (in district town, or city centre) user (login frequency)</td>
<td>Weekly ~ 1 hour</td>
</tr>
<tr>
<td>Village information centre user (login frequency)</td>
<td>Bi-weekly login</td>
</tr>
</tbody>
</table>
do not establish direct contact with all small farmers. They talk only to a small group of farmers to whom they offer information. They expect other community members to learn from this chosen group. Similar top-down information channel has been evolved in the information dissemination processes for social improvement of the community. Figure 2 shows a form of information channel.

In this top-down hierarchy, information on development issues does not flow directly to everyone in the community. The information flow is a multistep process. It flows first from the media, research team, and extension workers to opinion leaders or members of the information groups. From these small groups it flows to all the peasants and community members.

Numerous forms of ICT techniques exist that can support the communication of geographically dispersed communities. The article emphasized the dual nature of ICT, “which focuses attention on how information technology shapes human action through its provision of structural opportunities and constraints, while also recognizing that information technology is itself the product of human action and prior institutional properties” (Orlikowski & Robey, 1991, p. 148).

Activity of a typical information centre can be measured by the log in behavior of its clients. Usually, participants from city centres have better access to information than from village community centres. Similarly, city centre users have been found to login to the network (Internet) more frequently. Therefore, login frequencies of these two groups are different. Their behavior of login is shown in Table 1.

The degree of a social network gives an indication of how many entities a user interacts with. The average degree is calculated by:

\[
\frac{1}{n} \sum_{i=1}^{n} d(n) \]

where \(d(n)\) = number of edges incident on node \(i\), \(i = \) node as depicted in Figure 1.

Research studies indicate that degree of around unity in a typically medium network is acceptable, while degree of around 1.5 can be taken as medium interaction, and greater than 1.5 can be referred as heavy interactions.

Similarly, density of a virtual network can also need to be defined, to give the network the completeness.

\[
\text{Density} = \frac{e}{n(n-1)}
\]

where, \(e\) = number of edges in a network, and \(n\) = number of nodes.

The actual density values will decrease geometrically as the network increases.

Another indicator about the relationship among the member of a network can be derived, if the edges of the network (Figure 3) can be set in a symmetrical matrix such as:

\[
\begin{array}{ccccc}
  & a & b & c & d & e \\
 a & - & 1 & 1 & 1 & 1 \\
 b & 1 & - & 0 & 0 & 0 \\
 c & 1 & 0 & - & 0 & 0 \\
 d & 1 & 0 & 0 & - & 0 \\
 e & 1 & 0 & 0 & 0 & - \\
\end{array}
\]
while in Figure 3, a knows b, c, d, and e. But the relationship between b, c, d, and e may not be known.

Terms such as *global information highway*, *information revolution*, and *information economy* are used routinely to illustrate the profound role of information in modern societies. Yet it is easy to make unverified assumptions about the nature of benefits being generated in an “information society” (McConnell, 1996). However, measuring the social impact of a virtual network is not so easy and directly derived by conventional definitions. The impact may be defined as dependent on several attributes (± is used, as there may sometimes be a negative value);

\[
i_s = \text{social impact of a network} \\
\approx \text{inverse of the density value} \\
\pm \sum \text{variance (relationship value of adjacency matrix)} \\
\pm \sum \text{variance (degree of social network)} \\
\pm \sum \text{factors related to socioeconomy of a society.}
\]

**RETROSPECTIVES**

This section describes a few critical issues related to the importance of content, capacity building of the stakeholders, and a few characteristics of virtual networks.

**Importance of Local Content**

Knowledge is a complete product of human thought, action, and wisdom. Each culture contains a knowledge base from which its members receive understanding and an interpretation of the outside world. In building information content, it is important to start with the information and knowledge that exist within the community itself. Indigenous knowledge forms an important part of the community’s information and knowledge resources.

**External Information and Content**

Similarly, external information and knowledge sources are also vital because some human development problems are similar globally. By gaining access to information and knowledge sources that bring this experience to the community, members save valuable time, effort, and resources. They are able to learn and adapt ideas from the best practices of other grassroots communities in other countries.

**Raising the Capacity of Semiliterate and New Literate Members**

Feeding adequate information to this group is essential because they sometimes constitute about 30% or more of all the adult population of a community. They, therefore, have a big contribution to make in the development process.
Publication of a Community Newsletter

A newsletter to inform the community about what is going on, what is planned, and what could be the future model by information sharing and interchange activities. The newsletter could be issued every month, bimonthly, or quarterly. Copies of the newsletter can be put on notice boards of the community centres and other public places, with an electronic form saved in the Web server for Internet-connected participants and global viewers.

A successful virtual network should be:

- Consistent and visible to the link partners;
- Compelled to develop trust among participants;
- Clearly identifiable with common purposes in which the objectives of the linkage fit well with the mandates of the participating institutions;
- Accommodative to the willingness of participants to give up autonomy of action within the joint activity;
- Able to establish common understanding;
- Carefully managed so that the contributions of all participants are recognized;
- Capable of obtaining feedback between the strategic management team and the researchers;
- Containing linkages based on specific scientific initiatives rather than generic common interests to ensure direct benefits;
- Generating high-quality work with respect to the community; and
- Capable of providing good communication platform among participants and the outside world.

These networks should avoid:

- Expectations that signing a memorandum of understanding (MoU) will lead to all like-groups in establishing close working relationships and joint project activity;
- Impressions that the linkage may exclude collaboration with similar parties;
- Unclear agreement on the goals of the joint activity;
- Restrictive agreements that risk impeding natural extension and flexibility in the partnership;
- Processes that may lead to failed delivery of local benefits to the stakeholders; and
- Restricting decisions and activities in the partnership to reach all stakeholders/partners.

FUTURE ISSUES

Measuring actual social impact of a virtual network is a critical issue and needs further scientific and sociological studies and research. Performance behavior of communication mechanism can be improved by creating symmetry among the adjacency matrices through proper data calculation and analysis.

Centrality of the network (especially local centrality, in case of dispersed village information centres) is another important aspect of a virtual network, putting greater impact on social behavior of the participants. By calculating degree of interactions in every node, a graph can be generated, and the node with the highest degree may be taken as the center of the network. This concept may lead to a “star topology” concept, but this pattern may not seem to fit in real-life conditions during implementation period.

The exact nature of communication processes in virtual forms, their antecedents, and consequences are, of course, unknown as of yet. However, it is possible to glean some insight the rich body of literature on synchronous and asynchronous electronic organizational communication. Communication in the virtual organization certainly will become more electronically medi-
ated than in the past, and the vast set of empirical findings regarding mediated communication can foreshadow how communication will change (DeSanctis & Monge, 1998).

CONCLUSION

Economic growth is a poor indicator of development in this globalize world, because it does not reflect the actual distribution of wealth. It fails to show the impact of this development on other social segments, such as health, education, and gender equity.

An alternate development approach has come into being. This approach recognizes the importance of the well-being of all people. The name for this approach is people-centered development.

Among other social aspects of human development, the people-centered development approach should include the following issues in accumulating contents:

- Life expectancy and the health situation of the community
- Access to basic education and literacy of the community
- Gender equity
- Greater participation of community people in government decision processes

People-centered development effort also requires large amounts of information from the various development sectors to reach the grassroots level. For the community to take responsibility in the development process, it needs information on the following:

- Agriculture and livestock
- Environment
- General consumer commodities
- Primary healthcare facilities
- Education and literacy programs
- Leisure and entertainment
- Indigenous technologies and local knowledge
- Income generation
- Basics of demography (population, child mortality rate, social organizations in the locality, land disputes, and issues on human rights)

The role of information services in the human development approach (developing country concept) is very much different from the role of information in the advanced approach (developed country concept). In this aspect,

- Access to information should be for all groups in the community;
- Information should be treated as a tool for building self-reliance, social empowerment, civil society participations, and gender equity;
- Indigenous knowledge and locally generated content should be given priority; and
- Traditional channels of communication should be given preference.

In discussing information and knowledge, it is useful to start by observing that both are human creations (or social constructs). They are designed to explain and meet some of the challenges that individuals or groups face at a particular time and place. No one can fully understand the meaning of knowledge and information without recognizing that they can be a double-edged sword. They can be used to empower the individual and the group. They can also be used to continue relations of power and control. When a ruling group uses information and knowledge to control (dominate) people, those people are led to despair, powerlessness, and unsustainable lifestyles (Mchombu, 2002).

From a cumulative perspective, this research suggests that collaborative communication cannot be completely enforced via ICTs. ICTs may be essential but they are not adequate for the soci-
etal transformation promised by the proponents of virtual networking essence. The adjustment to an open information-sharing environment instigated by virtual networking can be seen as challenging to the community participants who are accustomed to a relatively reluctant working environment. Despite the many positive reports on virtual entrepreneurship, it is crucial to consider that intraorganizational communication is burdened with unrecognized intricacies, and added complexities of interorganizational relationships.

Because purely virtual entities are still atypical, the processes for developing virtual organizations and the eventual impacts of virtuality are still unknown to the researchers. Proponents of virtual networking exalt the benefits in terms of easy accessibility, greater adaptability, faster response time, and task specialization, while critics argue on potential downsides, such as broader divergence, decreased uniqueness, and higher probability of disastrous effects, including physical network disruptions.

REFERENCES


Mchombu, K. J. (2002). *Sharing knowledge for community development and transformation: A handbook*. Ottawa, Canada: DLR International, Canada with support from CIDA.


Rahman, H. (2000, September 14-17). Integration of adaptive technologies in building information infrastructure for rural based communities in Coastal Belt of Bangladesh. Paper presented at the *First Conference of the Association of Internet Researchers*, University of Kansas, Lawrence, KS.


KEY TERMS

Development Processes: These are processes for the improvement of society and community, through diversified means, creating demands and solutions at the appropriate dimension of necessity.

Information Dissemination: Content of any book, paper, map, machine-readable material, audiovisual, or other documentary material, regardless of physical form or characteristics, making them available to the public is a form of information dissemination. Means of dissemination may be through information resource databases, information portals, the Internet, or other available media.

Information Exchange: Putting ideas and concepts in the correct formats and get them circulated to other persons can be termed as information exchange. However, experiences and prior knowledge about the content are essential for making it become knowledge.

Social Network: It is a group of peers, subordinates, and interconnected people who provide general information about what is going on within the organization, ask specific advice on how to accomplish assignments, and obtain feedback about the consequences of different strategies.

Virtual Network: A form of network that provides virtual circuits and that is established by using the facilities of a real network. The term virtual network is used to refer to the appearance of a single, seamless network system.

Virtual Networking: It is a process of using the virtual network to gain access to specific resources, and exchange of information for mutual benefit reflecting an initial level of trust and commitment among participating organizations.
Chapter 6.2
Gender and Telework in Information Technology

Paula F. Saddler  
*Association of Certified Fraud Examiners, USA*

Donald D. Davis  
*Old Dominion University, USA*

Katherine A. Selgrade  
*Old Dominion University, USA*

Debra A. Major  
*Old Dominion University, USA*

INTRODUCTION AND BACKGROUND

Information technology (IT) work is often distributed geographically through practices such as teleworking. Telework lends itself well to IT workers because they work easily with information technology, which is required for telework, and because many IT jobs consist of knowledge work—the creation and analysis of symbols and ideas—which may be done anywhere and anytime.

Advances in information technology make distributed work possible. Globalization and the need for organization flexibility make distributed work necessary (Davis, 1995). Organizations distribute work to take advantage of scarce and inexpensive talent, enhance innovation and product design, and to reduce real estate costs, development time, and labor costs. Workers choose distributed work to balance work and life demands, reduce commuting time, accommodate disabilities, and take advantage of distant opportunities. Telework, a form of distributed work first described by Nilles (1975), has established itself throughout the United States. We discuss telework trends and provide some data describing teleworkers in IT professions in the United States.

Four forms of telework are commonly used (see Key Terms; Bailey & Kurland, 2002; Kurland &
Most teleworkers use a combination of these forms, although home-based telework is most prevalent (Davis & Polonko, 2001). Each form of telework is practiced for different reasons and produces different work experiences and outcomes (Bailey & Kurland, 2002; Davis & Polonko, 2003; Helling, 2000).

A national survey of telework practices in the United States was conducted in 2001 under sponsorship of the International Telework Association and Council (ITAC) and AT&T (Davis & Polonko, 2001). The sample was stratified to represent all U.S. households and was diverse with respect to gender, ethnicity, occupation, organization size, and industry. Results showed that there are approximately twenty-eight million teleworkers in the U.S. Compared to nonteleworkers, teleworkers are significantly more likely to be from the Northeast and West, male (54% of teleworkers), have higher education and income, work in professional/managerial occupations, work in industries such as construction, professional/scientific/technical services, health care/social assistance, and work in very small and very large organizations. There were no significant differences in telework practice for marital status, race/ethnicity, and age.

Davis and Polonko (2001) report several findings concerning telework and its impact on ability to balance work and family demands that are relevant to IT professionals. Teleworkers, especially those who work at home, are more likely than nonteleworkers to report working longer hours and that the demands of their personal life take time away from their work. They also report more difficulty relaxing at home when doing non-work related activities. However, teleworkers are also more likely to state that their family members and friends report less unhappiness with the amount of time they spend working at home, report that their home life is less likely to prevent them from working the amount of time they want to work at home, and that working at home interferes less with other activities in their personal life.

These differences suggest that teleworkers may be better able to manage the time they spend working at home compared to nonteleworkers. While teleworkers may experience some personal costs, such as difficulty relaxing, they experience significant benefits compared to nonteleworkers in terms of less interference between work and family roles. Women may particularly benefit from telework because they are often responsible for childcare.

Research examining women teleworkers shows mixed benefits. Women teleworkers may experience less stress because they have more control over their time, do less commuting, experience fewer distractions, and are available to their families during the day (Kraut, 1988; Olson & Primp, 1984). On the other hand, women teleworkers may also experience greater stress due to reduced separation between work and family obligations resulting from bringing the workplace into the home (Olson & Primp, 1984). All teleworkers report working more hours (Davis & Polonko, 2001; Mirchandani, 1998).

Men and women adapt differently to work at home. For example, the time saved by reduced commuting gets used differently; women tend to use this time to do household chores, whereas men use this time to do more job-related work (Steward, 2000). Both men and women maintain a separation between work and family, but they do so for different reasons. Mirchandani (1999) reports that most women state that they separate work and family to reduce the stress of simultaneously attending to work and family obligations, whereas no men report such reasoning. She adds, “For women, the home is not a place of nonwork, but rather another workplace” (p. 92). Women teleworkers, especially those with children, may experience more difficulty in balancing work and family life than men (and women without children). Yet women claim to choose to telework as a means for balancing work and family obligations (Beasley, Lomo-David, & Seubert, 2001).
TELEWORK IN IT ORGANIZATIONS

Telework may be easier for IT workers due to the knowledge content of their work and their facility with using technology, although research results are mixed. A study of IBM teleworkers and non-teleworkers showed no difference in work family balance (Hill, Hawkins, & Miller, 1996; Hill, Miller, Weiner, & Colihan, 1999). In participant interviews, many claimed that teleworking helped them fulfill family obligations and strengthen family relationships, whereas others felt that teleworking blurred the boundary between work and family life and therefore made it more difficult to balance work and family demands.

Telework affects different aspects of work-family conflict. Selgrade and Davis (2005) studied a sample of teleworkers that included IT workers as well as workers from other professions from eleven organizations throughout the United States. They found that the effects of work interfering with family (WIF) were different than the effects of family interfering with work (FIW). Results showed that WIF, but not FIW, was negatively related to telework satisfaction and organization commitment and positively related to turnover intentions. Moreover, WIF increased with telework duration, whereas FIW increased with the number of children (Selgrade, 2004).

We report the following findings from a study of IT workers in eleven organizations. Major, Davis, Sanchez-Hucles, Germano, and Mann (2006) describe the research in more detail. We report here differences between IT teleworkers and nonteleworkers. All differences are statistically significant except where noted.

Forms of Telework

We found that 39.8% of IT workers report teleworking (less than one day per week to five days per week), mostly from home. Of those teleworkers who reported working at home, the majority of them do so less than one day per week (78.3%), whereas 21.7% report working from home one day per week or more. Of those teleworkers who reported working at telework centers or satellite offices, the majority of them do this less than one day per week (59.1%), whereas the remainder (40.9%) report teleworking one day per week or more. Finally, of those teleworkers who reported working on the road, 47.2% report doing this one day per week or more.

Demographic Profile of Teleworkers

Gender

Females comprise 39% of the total sample; only about 1/3 (36.3%) of them telework (see Table 1). Males make up 61% of the total sample, nearly 1/2 (48.3%) of them telework. Consistent with national data (Davis & Polonko, 2001), male IT workers are more likely to telework than women.

Race

Whites were slightly more likely to telework, but there were no significant racial differences between teleworkers and nonteleworkers. The finding of racial balance in teleworking is consistent with national patterns (Davis & Polonko, 2001).

Relationship Status

We examined whether IT teleworkers tend to be single or in a committed relationship. We define “single” as those individuals who reported being single, divorced, or widowed and “committed” as those individuals who reported being married or living with their partner. Teleworkers and nonteleworkers were similar in their relationship status. These results mirror national data (Davis & Polonko, 2001).
Number of Children

Teleworkers and nonteleworkers were similar in the number of children they had. When examining only teleworkers, we found that the gender breakdown mirrored that of the overall sample; male teleworkers tended to have slightly more children ($M = 1.29$) than did female teleworkers ($M = 1.02$).

Highest Degree and Salary

Consistent with national patterns (Davis & Polonko, 2001), teleworkers had a higher educational level and a higher salary than nonteleworkers. IT workers in our sample were highly educated; most (71.7%) had at least a bachelor’s degree. More than three-quarters of teleworkers (76.2%) had at least a bachelor’s degree, while this was true for just over 2/3 of non-teleworkers (68.5%). This difference was not statistically significant. While males and females did not differ in terms of their highest earned degree, they did differ in terms of their salary; not surprisingly, males earned significantly more than females. When examining teleworkers only, men still earned higher salaries than women.

IT Position

Participants were asked to indicate their current IT position by selecting one of the following categories: conceptualizer, developer, modifier/extend, supporter/tender (see Key Terms). The breakdown was very different in terms of teleworkers vs. nonteleworkers. Proportionally more teleworkers were conceptualizers; proportionally more nonteleworkers worked in supporter/tender positions. These findings are not surprising because supporter/tender positions require IT workers to be in close proximity to the equipment and people they are responsible for supporting, whereas conceptualizer jobs do not. Supporter/tender positions may not allow the geographic and scheduling flexibility required of telework.

Elder Care

We were interested in whether or not teleworkers more than nonteleworkers chose telework because they were responsible for the care of an elder adult and needed the freedom to work from home. There were no significant differences in elder care between teleworkers and nonteleworkers.

Individual vs. Team Work

We asked participants whether they worked alone or as part of a team. We thought that the benefit of telework might be restricted to those who do not work with others. We found that teleworkers and nonteleworkers were similar in emphasis on individual vs. teamwork; the gender breakdown within this variable was similar as well. This finding lends support to the notion that even individuals who work in teams can telework.

Total Hours Worked

We examined the total number of hours worked per week by IT workers. Previous research suggests that one of the disadvantages of telework is that teleworkers tend to work more than nonteleworkers (Davis & Polonko, 2001; Mirchandani, 1998). The findings from the current sample support previous research; teleworkers in our sample worked significantly more hours per week ($M = 48.55$) than did nonteleworkers ($M = 45.25$). In addition, for the entire sample, males worked significantly more hours per week ($M = 47.52$) than did females ($M = 45.42$). However, when we examined the teleworker and nonteleworker samples independently, the gender results were different. In the teleworker sample, males worked significantly more hours per week ($M = 49.37$) than did females ($M = 46.92$). In the non-teleworker sample, there was no significant difference between males ($M = 45.80$) and females ($M = 44.51$) in the number of hours worked per week. These results are consistent with previous
research that suggests that men and women use the extra time they save through teleworking in different ways. Males tend to use this time to do more job-related work, hence their longer working hours; whereas females may use this time to do more family-related work such as household chores (Steward, 2000).

CONCLUSION AND FUTURE TRENDS

Telework is a growing response to the need of organizations to reduce costs, enhance flexibility, and attract and retain workers with valuable skills. It is one way to create virtual teams and virtual organizations. Inexpensive information and telecommunications technologies and increasing global competition will accelerate this trend for IT organizations (Davis, 1995). IT workers, because they engage in knowledge work and are skillful at using information technology, are particularly well suited for telework.

Researchers are just discovering the consequences of telework and how to manage it effectively. Much telework research has been limited by reliance upon small convenience samples, poor measurement, inadequate statistical analysis of results, and inadequate grounding in organization theory (Bailey & Kurland, 2002). Management practice has lagged as well, with little understanding of how to alter organization structures and processes to make telework practice successful (International Telework Association and Council, 2000). There is a great need for well-designed research studies of telework in IT.

Women in many professions throughout the United States participate in telework. Women in our sample of IT workers participate extensively in telework as well, but when compared to men, they participate at a lower rate in comparison to national practice (IT teleworkers in our sample: 67% male, 33% female; national sample in U.S.: 54% male, 46% female; Davis & Polonko, 2001). This difference may be unique to the organizations in our sample. This difference also may be due to institutional barriers such as resistance to remote work among IT managers and work design barriers such as the need to tend to centrally located equipment around the clock. Future research should continue to explore gender differences in telework among IT workers and identify possible barriers and enablers. These results may then be used to help IT organizations and IT workers increase flexibility by distributing IT work away from the traditional office. This flexibility will enhance efforts to balance work and family needs as well as organizational competitiveness.

REFERENCES


KEY TERMS

Conceptualizer: Someone who conceives of and sketches out the basic nature of a computer system. Job titles include: product designer, research engineer, systems analyst, computer science researcher, requirements analyst, and system architect.

Developer: Someone who works on specifying, designing, constructing and testing information technology. Job titles include: system designer, programmer, software engineer, tester, computer engineer, microprocessor designer, and chip designer.

Distributed Work: Work that is distributed across geography and time. Distributed work may take several forms, (e.g., telework, telemedicine, virtual teams, and virtual organizations).

Knowledge Work: Work that consists of the creation and manipulation of ideas and symbols. Knowledge work is common in IT and is ideally suited for distributed work because it is easily digitized and may be done anywhere and anytime.
Modifier/Extender: Someone who modifies or adds on to already created information technology. Job titles include: maintenance programmer, programmer, software engineer, computer engineer, and database administrator.

Supporter/Tender: Someone who delivers, installs, operates, maintains or repairs information technology. Job titles include: system consultant, customer support specialist, help desk specialist, network installer, and network administrator.

Telework: A type of distributed work in which employees work at locations other than the traditional workplace. Major types of telework include: (1) home-based telework—working from home; (2) satellite offices—employees from the same company work at the same location closer to their home, often in suburbs or exurbs; (3) telework centers—similar to a satellite office except that the telework center hosts employees from different organizations; (4) mobile work—working at multiple locations away from the office (e.g., at client work sites, while traveling).

Virtual Organization: A form of organization design in which organization structures and functions are distributed across time and geography.

Virtual Team: A form of work design in which members of the same work group are distributed across time and geography; usually compared to traditional collocated teams, where all team members are located at the same site.

ENDNOTE

1 This material is based upon work supported by the National Science Foundation under grant no. 0204430.

This work was previously published in Encyclopedia of Gender and Information Technology, edited by E. Trauth, pp. 460-466, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global)
Chapter 6.3
Managerial Careers, Gender, and Information Technology Field

Iiris Aaltio
Lappeenranta University of Technology, Finland

INTRODUCTION

Careers are organizational and institutional, and they have know-how-based contexts. Managerial careers from a gender perspective, gendered “blind spots” in organizations and the invisibility of women in management have been an object of study since the 1970s. Gender is a part of socially constructed individual identity. Gendered identities in organizations are defined and redefined in relationships as people become socially constructed through work groups, teams and interactions. Because of this social construction, femininity and masculinity grow into human behavior and outlook. Understanding gender as an activity and a term in the making (Calás & Smircich, 1996), it is a constitution of an activity, even when institutions appear to see woman and man as a stable distinction (Korvajärvi, 1998). Beyond work-life and organizations, there are multiple institutional and gendered structures. The information technology (IT) industry and companies are also an institutional construction with gendered dimensions, and they also participate on the creation of femininity and masculinity.

Career can be seen as a conceptual artefact that reflects a culture and rhetorical context in its use. It is a kind of window to a network of values, institutions and functions, where actual careers are made. Usually, the formal organization is based on neutrality and equality, but a closer look reveals the deeper social structures that make it different to women and men. There is a concept of an abstract and neutral worker, and this worker is supposed to be highly competent, work-oriented and available, committed to work-life without any knit to private life. These characteristics support a good career climb in an organizational hierarchy, and many of these characteristics better suit men than women (Metcalfe & Altman, 2001). For instance, home responsibilities make often working hours less flexible for women than men. The notion of an essential person with no gender characteristics does not recognize these issues, whereas taking gender as a research topic shows that work-life as a context differs between women and men.
BACKGROUND

Managerial Positions within the IT Industry

Organizational structures, including managerial positions, are gendered by nature. Overall, there is a high degree of vertical segregation, which means that there are few women in managerial positions compared to men (Acker, 1992). According to the United Nations’ *World’s Women 2000* report, women’s share of the administrative and managerial labor force is less than 30% in all regions of the world. This is true also in Nordic countries, where the participation of women in work life is almost 70% and has a long tradition. Women also hold only 1% to 5% of all the top executive positions (Wirth, 2001), and the numbers seem to change very slowly. In the European Union countries, women’s share has barely changed since the early 1990s, and has remained at a less than 5% level (Davidson & Burke, 2000). This division of managerial top positions is called glass-ceiling phenomena, and it exists world wide (Powell, 1999): “The higher the managerial position the fewer the women.” As a result of this, women and the highest economic power become separated.

Taking a closer look at the numbers, the least amount of women in top positions are found in male dominated areas, such as heavy industry and construction business, where the amount of female leaders is less than 10%. IT is also a male-dominated field. There are few female directors in an organization that employs mostly men (Kauppinen & Aaltio-Marjosola, 2003). The number of female managers has increased slowly. In many fields, like IT, it is still low (Ahuja, 2002). Women’s and men’s work in organizations also differ from each other by nature; that means women and men end up doing different kinds of work horizontally.

Current statistics indicate that women account for about 25% of technology workers in the European workforce and about 20% of those in the United States’ (U.S.), and that there looks to be a polarization in the type of work women and men do. The majority of women are employed in routine and specialist work, like clerks, while men are engaged in analytical and managerial activities. In the studies, overall 10% of males and only 3% of females within IT had achieved senior managerial positions (Ahuja, 2002). Salary gaps for women and glass-ceiling perceptions are reported as well in this research. Despite that the IT profession has grown in recent years, there remains a gender imbalance and, in some cases, even evidence of a decline in female workforce numbers (Ahuja, 2002).

Managers and leaders have identities that become constructed within special circumstances, and IT constitutes a particular background for identities to grow in. As stated by Davis (1995), organizations and their activities are cultural constructs arising from the masculine vision of the world, and IT’s close connections to the male-dominated technology field and its high numbers in male participation makes its connection to masculinity evident. The glass ceiling in the IT field might even be stronger that in others, because there is evidence that women there tend to be stereotyped as staff, the ones who don’t take risks, rather than “line” people; whereas men are the innovators and designers (D’Agnostico, 2003; Russell, 2004). This results on men’s career outcomes including higher managerial positions.

The segregation of work is based on the classical stereotypes of women’s and men’s behavior and orientations. Men are oriented towards technical and industrial work, whereas women are engaged in occupations where one needs caring ability and social integration, such as teachers and nurses. Ideals for men’s and women’s work differ from each other and carry stereotypes (Aaltio, 2002). Women and men are easily valued differently because of their gender. In society, there are different places for women and men, and this holds both in families and in work organizations. Men
historically relate to the public and women to the private spheres of life (Acker, 1992). By extending their roles and breaking into public institutions, women challenge the prevalent male ideals and bring private issues into public and institutional spheres. Career opportunities mean different things for men and women and, therefore, challenge them differently.

It is also notable that the relationship between gender identity and participation in the IT profession is not the same for all women, nor is it based on monolithic values that are the same for all women. Each individual woman experiences societal influences differently and brings her personality and characteristics into the field (Trauth, 2002), even if women’s institutional backgrounds are similar.

**WOMEN’S CAREERS WITHIN THE IT FIELD**

Discourses of Career: Interviews from IT Managers

In the study, five managers were interviewed, two of them women and three men. They all work in a large, successful IT industry company in Finland, are young or early middle-aged, and are labeled as high achievers with a promising future in the company. In the interviews, it was asked how they see their managerial career development up until now and in the future. The data gathered was narrative by nature, because a lot of space was given to open talk around the topics raised at the interview (Aaltio-Marjosola, 2002).

As one of the managers describes: “Professional development is my area, and I am globally responsible [for] this part of the business. The culture here in the company is, however, where I start from. In my close network, we have a team of 10 people, but globally, it is a big circle, of course.” As seen in the citation, instead of manager-subordinate relationships, she emphasizes the close network and teamwork orientation. Another manager illustrates his job with no clear subordinate relationships: “I came here to coach a team with a few people and to coordinate things. This is more or less process management; I have to think first and then make things simple and concrete, communicate them to the others. I work closely with the head of the company in some projects, in addition. In fact, I do not see that there is a clear organization where I am a leader.”

Further, one of the IT managers describes her job: “I am a personnel consult, even if I also take care of the unit’s personnel management, and my work is multiple of things. For instance, I also do training and lecturing in our inner training seminars.” In spite of the high organizational position, she sees her job to be consultative, including support of teams.

It is also typical to describe career development that includes multiple paths, work in many organizations and rapid movements during the career. One of the managers described his work: “I find that my career is more or less a bumpy road, not going upward all the time. I have started two times from a beginning in my job and found that you have to take your space in every new job no matter what is your position when you start. You have to find your fit to the values of the enterprise, feel they are near to yours, to get the right start.” Still another manager describes his career development: “My job is changing. I will take care of the process development here, coordinate and harmonize things. I have to move as rapidly as the whole company does, this is the way it comes here. The number of my subordinates will be like 50 or 100 altogether, I am not quite sure how many they are, in fact.”

Career, in the descriptions of these managers, is advancement of abilities, seeing oneself as an important and integral person in a certain close network, and giving good support for subordinates. None of the interviewed managers started their description of their work from the organizational ladder or other kind of embodiment of hierarchy.
Managerial Careers, Gender, and Information Technology Field

This shows the IT industry’s development, being at the top of companies that themselves develop work practices and communication styles that break traditional ones with hierarchy and high formal authority. Much of the work is based on projects and close teamwork. Both men and women adapt into this if they seek to be successful in the company and get advancement.

Managerial Careers of Women within IT

Today’s careers are based on a variety of choices; careers are “boundaryless” due to multiple routes and individual choices. To predict career development is more difficult than it used to be. Individuals’ knowledge and skills give them good possibilities to move from one job to another, and organizations seldom serve for a life-time basis in a person’s career development. Flexibility and adaptability of careers have increased, and diverse assignments are common instead of a clear and one-path kind of career development (Arnold, 2001; Storey, 2000). Work will become rich from difference and requires abilities and motivation typical for women, who have always combined many roles, private and public, and thus learned flexibility.

For instance, Ahuja (2002) suggests that there are barriers to female managers’ careers within IT, and that their minor positioning in the IT industry is due to an “old boys” network. There is a large pool of qualified and experienced male professionals, whereas the pool of women is still few. This counts for a lack of female role models and mentors for the younger generations, and discriminatory practices remain. There is also evidence that women do not see IT as an attractive option, and even if they do, they are technically less equipped to come over. There are few managerial role models for women in IT industry companies, and women would benefit from career consulting.

Himanen (2001) argues that computer hackers will become heroes of the information society, and the heroes of IT are and will in the future be men, as seen in, for example, Silicon Valley, where there are very few women in higher-level positions. However, men also represent a variety rather than a unity. Taking the fluid notion of gender, we can see that masculinities are carried by organization cultures and are not unitary. In the changing work life, advancement also requires new skills and attitudes from men, not only for women as a unitary category.

Also, family-work contrast barriers were reported. A survey (Prencipe, 2001) outlined attitudes of IT women leaders, half of them in senior positions. The study showed a positive personal valuation of one’s work, but also that balancing personal and professional life was still complex for most of them. Reported stress was due to working late, constant change and sometimes reported discrimination in a social environment consisting of mainly young, often male colleagues (Prencipe, 2001). When managerial career advancement starts from lower professional levels, factors like role stress tend to decrease job satisfaction, resulting in lower career outcomes of women (Igbaria & Chidambaram, 1997). In addition, at the institutional level, where changes are very slow, women and technology are still a combination that arouses suspicion.

FUTURE TRENDS AND NEEDS OF RESEARCH

Challenges of IT women managers’ career development are in line with the results of earlier gender studies. In addition, the attractiveness of the company cultures for women managers is also questionable, because of the cultural codes that might be gender-biased, repulsive and more favorable for men to follow. These cultural codes need to be studied and better understood.


As a part of high technology, IT has taken the world by storm and is changing the way businesses learn, as well as the nature and characteristics of work. The implementation of IT within and across organizations is reducing the importance of hierarchy and command-and-control authority systems that structure power within them. High technology changes the traditional managerial and communicational style from vertical to horizontal (Zeleny, 1990). In high technology-oriented industries, power is connected with expertise, which may break down the traditional hierarchy in an organization (e.g., Gunz, Evans, & Jalland, 2002). Research is needed on the special nature of work in IT companies, in order to understand which kind of management they really need, in terms of good results and higher well-being of the work force.

IT enterprises are themselves examples of “new” ways of management and doing work. It has been stated that the IT industry is in transition between the old sense of identity and the new one (Colwill & Townsend, 1999). As argued, the old culture was directed towards providing the answers, not at meeting the needs of the users. Working in intense networks is increasing and good communication skills are highly needed in the future. Gender-biased cultural expectations and gender stereotypes still work as a barrier for female managers’ careers in IT. However, the changing nature of IT work and its prevalent practices make it open to diversity of expertise and values. Both women and men can benefit from these developments if the IT field itself learns gender-equality and tolerance.

The future’s industries also compete to commit good expertise, and women are a resource. Multiple skills and their good management are needed, and new solutions to old problems will be searched. This also will advance women’s managerial careers.

CONCLUSION

IT-field companies appreciate technology development. They have to learn understanding of social and cultural aspects that, however, are the true background for their development in the future. The field should attract both capable women and men. While there are a few managerial role models for women in the IT industry, women involved would benefit from career consulting and mentoring programs (Ahuja, 2002). The use of tele-work and e-leadership might also positively affect women in the industry, because they support combining family and work life (Avolio, Kahai, & Dodge, 2000; Beasley, Ewuuk, & Seubert, 2001). This kind of supportive activities may attract more women to the field and advance women’s managerial careers. Organizational cultures that grow towards multiplicity of values and a variety of femininities and masculinities may be better work environments, especially for women but also for men in the future.

REFERENCES


**KEY TERMS**

**Career:** Consists of the sequential choices made by a person. It is a developmental process of professional identity and personality.

**Culture:** A human process of constructing shared meaning that goes on all the time and is based on human unique capacity for self- and other-consciousness.

**Doing Gender:** Concerned with the constitution of gender through interaction, and makes gender a consequence of an ongoing activity.

**Gender:** An integral part of socially constructed individual identity that constitutes and restructures a multitude of cultural and social phenomena.

**Gender Bias:** To generalize essential individual behavior without gender dimensions.

**Gender Stereotypes:** Refer to the traits and behaviors believed to occur with differential frequency in the two gender groups of women and men.

**Glass Ceiling:** A metaphor that describes the tendency of women get excluded from top management in organizations in spite of their occupation of middle-management positions.

**Horizontal Gender Segregation:** Refers to the structure of the workforce that tends to become separated into women’s and men’s work.

**Organizational Culture:** A collective construction of practices, meanings and expressions that can be seen developed in interaction within organizational social spheres.

**Vertical Gender Segregation:** Refers to the structure of organizations that position women on lower managerial levels compared to their male counterparts.
Chapter 6.4
Social Networking and Knowledge Transfer in Collaborative Product Development

Katariina Ala-Rämi
University of Oulu, Finland

ABSTRACT

Software product development requires connecting of specialized information and know-how. Therefore, planning and production are widely done in networked projects. This chapter studies inter-firm collaboration to explore the role of face-to-face meetings and information and communication technology in knowledge transfer in product development by using case studies of software companies in Oulu, Finland. Clusters are important in forming trust and mutual understanding. Thus, a compact city region and effective social networks are beneficial for inter-firm collaboration. However, information and communication technology is very valuable for routine communication and documentation, but also in distance collaboration. Geographical distance affects inter-firm collaboration in terms of practical arrangements, such as working hours, yet cultural differences bring additional challenges. This chapter considers and scrutinizes social networking in collaborative product development, and the role of clusters and information and communication technology in such development, as among the key success factors.

INTRODUCTION

In an increasingly globalized economy, the competitiveness of an urban region depends on how well firms adjust to increasing competition and growing complexity. In knowledge-based sectors, social networks of actors are considered vitally important in keeping a firm’s information and know-how up to date. Besides being an individual process, learning is also a social process in a specific social context (Wenger, 1998). The functional urban region of Oulu (FUR Oulu),
with a population of 210,000, is regarded as one of the most successful information and communication technology (ICT)-based economic regions in northern Europe (Oinas & Lagendijk, 2005, p. 324). Strategic decisions at the University of Oulu and government investments were the starting point for the growth of Nokia’s mobile phone industry, which has with its subcontractors formed the foundation of the technology cluster in the Oulu region (Donnelly & Hyry, 2004, pp. 136-138; Männistö, 2000, p. 84). In spite of the success, the ‘Oulu Phenomenon’ could not be understood without Nokia’s development (Männistö, 2002, pp. 2002); additionally a small group of skillful people with strong mutual trust and a shared goal have been in a crucial role in that development (Donnelly & Hyry, 2004, pp. 135-136; Tervo, 2004, p. 86).

The high technology cluster in FUR Oulu consists of five different branches, of which information technology is covering over a half of the enterprises. Other technologies are multimedia, health and wellness, biotechnology, and environmental technology (Donnelly & Hyry, 2004, pp. 135-136). In the region, information technology covers mostly telecommunications and software. However, classification of different technology cluster is complicated, since a remarkable amount of telecommunication is assessed to belonging to a software cluster (Männistö, 2002, pp. 199). The software cluster, with strong connections to telecommunication and also wellness sectors, is expected to develop rapidly in FUR Oulu (Männistö, 2002, pp. 184). Moreover, software products are complex products that require many combinations of specialized information and know-how. Therefore, planning and production are more likely to take place in networked projects (Miettinen et al., 2006, p. 26).

Nonetheless, the region meets new challenges along with global development, as Nokia among other long-standing successful firms is increasingly moving its production abroad, mainly to the Asian countries, where production expenses are lower and human resources are abundant. Nokia still plays an important role in investment in research and development (R&D), but there is a significant increase in the investments of small companies (Lassila et al., 2006, p. 31). The software industry has grown rapidly also other parts of the Finland, and the sector is considered a future possibility for economic growth (Finnish Software Business Cluster, 2006). Software production does not require transporting of heavy materials; the crucial issues are skilled labor and good infrastructures, so in principal the sector is well suited for a peripheral high-cost region such as northern Finland. By peripheral region, I mean in this case its physical location as the northernmost inhabited region far away from the main market areas, and even though Oulu is sixth largest city (130,000 inhabitants) in Finland, it is a small place globally. Moreover, northern Finland outside FUR Oulu is very sparsely populated (the population density of the whole region is 4.6 inhabitants per square kilometer). However, per capita investment in R&D in FUR Oulu is the highest in Finland.

Technopolis Linnanmaa, next to the University of Oulu, is the oldest technology center in the Nordic countries with over 100 high-technology firms, 4,000 employees, but also providing a good infrastructure and meeting facilities (Jauhiainen, 2006, pp. 1413). The idea of the Technopolis concept is to build a supportive high-technology business environment by providing business and development services in same premises (Launonen, 2006). Besides wireless Internet connections, these services include all kinds of practical support for business activities, from cleaning and copying to tailored packages of more demanding clerical and personal assistant services (Technopolis, 2007). Nowadays Technopolis has several buildings in FUR Oulu, located in strategically important places such as next to the University of Oulu, the city center, or close to airport (see Figure 1). Furthermore, the Technopolis has
premises in four other important cities in Finland and most recently started in St. Petersburg, so the Technopolis group is the largest sector in Europe with its 950 companies (Launonen, 2006).

The idea behind the Technopolis concept is the collection of so called ‘critical mass’ in the same premises in order to develop growth clusters by providing technical infrastructure and possibility to meet and network, while being located in same building. Being located in close proximity is widely seen as an advantage for collaboration and networking (Howells, 2000; Torre & Rallet, 2005). This study explores the limitations of distant collaboration and the city’s urban scale and diversity in developing growth clusters. It uses a specific case as an example to explore knowledge transfer in product development. It studies social networking in collaborative product development, and the role of physical and temporary clusters and ICT in that development.

BACKGROUND

To create new knowledge and innovation faster than competitors, interaction with local and global actors is vitally important (Asheim & Herstad, 2003, p. 3). Technological development is diminishing limitations in communication, so that factors other than physical distances are increasingly decisive in collaboration and networking. The knowledge contained in a firm consists of the people working in the firm and their ability to achieve new information and knowledge. Communication between individuals, also with people outside the firm, plays an important role in staying up to date in local and global development and know-how. Geographical proximity makes networking and communication between individuals smoother (Howells, 2000, pp. 53-55; Nachuum & Keeble, 2003). The most important way to find needed know-how or capabilities is to
use existing social networks. Another is professional gatherings, which Maskell et al. (2006, p. 999) call ‘temporary clusters’. Cafés, bars, and other informal meeting places can act as temporary clusters for exchanging information, but they are more random than organized meetings.

Nevertheless, even though geographical proximity is an advantage for networking and knowledge transfer between enterprises, it is not an automatic guarantee of or a compulsory condition for networking (Bathelt, 2005, pp. 108-110; Boschma, 2005, p. 63). Furthermore, geographical proximity is not relevant if there are no shared interests and goals, created in a spatial cluster with complementary or similar capabilities. Because of the above-mentioned, proximity is seen here as being relative, suggesting that there are different dimensions of proximity that are required in collaboration, such as organizational, temporal, cultural, cognitive, social, and institutional proximity, not just geographical proximity (Boschma, 2005; Oinas & Lagendijk, 2005, p. 311). Depending on the situation, different types of proximity have to be taken into consideration along with geographical proximity (Aslesen & Jakobsen, 2007, p. 191; Torre & Rallet, 2005, p. 49).

ICT supports the activities of enterprises by means of easy and fast modes of communication. ICT has also made information more ‘ubiquitous’ by accelerating the spread of new ideas, information, and innovations. Consequently, ICT has a central role in knowledge-based economy, in which knowledge generation and exploitation play a predominant part (Roberts, 2000, p. 429). As the information is more easily reached by everybody, also competition between firms has increased, creating pressure to produce new innovations and products faster. However, not all knowledge is ubiquitous and easily expressed with words, as Polanyi’s famous concept of ‘tacit knowledge’ suggests. Some knowledge is even subconscious and tied to physical skills. Some capabilities are built up over time in certain regions, such as interpersonal connections and a particular pattern of inter-firm networking, and not possible to transfer in other regions (Lawson & Lorenz, 1999, p. 310).

Roberts (2000, p. 429) expresses that ICTs “favor the transfer of knowledge that can be codified and reduced to data.” However, ICTs increase information about sources of knowledge and communication between distant actors in the same field, especially when sharing a common social and cultural framework, and thus is promoting R&D collaboration (Roberts, 2000, pp. 435-436). It has been argued that especially new tacit knowledge can only be diffused via face-to-face interaction (Andersson & Karlsson, 2004, p. 284). Cross-cultural limitations can also result in unexpected difficulties, as tacit knowledge is embedded in a social and cultural context (Roberts, 2000, p. 431). To make it possible to reach an agreement on the distribution of work and the rules of communication, virtual projects should be started with face-to-face meetings (Pyörä, 2005, p. 50). Furthermore, also codified knowledge is context-dependent, that is, in terms of common language or professional background (Lawson & Lorenz, 1999, p. 309; Roberts, 2000, p. 431).

The theoretical framework of this chapter is based on two approaches. First of all, social networks are important in forming and dispersing networks. Especially informal relationships are vital in this phase. Secondly, activities like product development and knowledge transfer take place within the community of practice (CoP), which is defined as a group of people with shared interests and aims (Hislop, 2004, p. 38). The concept of social network refers to the idea that social relationships consist of nodes (individual actors within networks) and ties (relationships between actors). Strong ties are more likely to exist in informal contacts with family members and friends sharing the same interest, knowledge, and trust. These relationships have special importance for networking. Formal contacts that mostly exist with clients, suppliers, competitors, and workmates are weaker, but they bring new ideas and opportuni-
ties to the network (Ellis, 2000, pp. 445-446). For knowledge spillovers, informal linkages between firms, such as networks of small firms, are important (Dahl & Petersen, 2003, p. 5).

The idea of the CoP concentrates on social learning, which requires that people have a common interest in some subject or problem and that they collaborate over an extended period to share ideas, find solutions, and build innovations. Wenger et al. (1999, p. 4) state:

...these people don’t necessarily work together every day, but they meet because they find value in their interactions. As they spend time together, they typically share information, insights, and advice. They help each other solve problems. They discuss their situations, their aspirations, and their needs.

In different CoPs there are not only differences between knowledge and common interest, but there can also be different assumptions and interpretative frameworks, leading to a situation where there is no understanding of the knowledge of the other community (Hislop, 2004, p. 40).

This study uses a specific case to explore the networks of software companies and knowledge transfer in product development. What role does the city’s urban scale and diversity have on the development of growth clusters? What knowledge is transferred by using ICT? When are face-to-face contacts necessary? How does the distance between collaborative partners affect product development?

FINNISH INNOVATION SYSTEM AS A PROMOTER OF SOFTWARE INDUSTRY

A national innovation policy has been implemented in Finland since the beginning of the 1990s. Nowadays, the Finnish innovation policy is based on a national innovation system and local innovation systems (Lievonen & Lemola, 2004, p. 40; Suorsa, 2006, p. 30). The Finnish innovation system is considered internationally competitive (Finnsight2015, 2006, p. 8), but vulnerable because of its small size which leads to limited resources and funding of research. Finland’s physical location in the periphery of Europe creates additional challenges; the population and economic activities are mostly concentrated in a very small region in the southern and southwestern parts of the country. Because of the limited resources, there has been a focus on a few fields and regions (growth poles) with enough resources and quality. The Finnish innovation system emphasizes knowledge, learning, know-how, research, and innovation for the competitiveness of Finland, but also the whole society (Suorsa, 2006, p. 30).

The development of the ICT industry and growth in labor productivity were essential to Finland’s rapid recovery in the 1990s from one of the deepest depressions of the country. However, Finland’s economy is facing challenges in regional development. There is strong dependency on the ICT sector, an internationally small number of small and medium-sized enterprises (SMEs) and an aging population (OECD, 2005, p. 7). Investments in R&D have gradually increased in Finland, being almost 5.3 million euros in 2004, which is 3.5% of the GDP. Enterprises’ share of R&D has increased up to 70%, and higher education constitutes up to 20%. Even though the investment of the public sector has increased in terms of the number of euros, its share has declined to 10% (Statistic Finland, 2006). Nokia still has an important role, providing almost half of enterprises’ R&D investments (Suorsa, 2006, p. 33).

As mentioned before, the Finnish innovation policy aims to support successful industries in regions with good economic growth possibilities (Lievonen & Lemola 2004, p. 55). The Centers of Expertise Program is the main instrument of Finnish regional innovation policy. The idea is to support networking of enterprises with research...
Social Networking and Knowledge Transfer in Collaborative Product Development

universities and universities of applied sciences in the region, public sector entities, and business support services in specific fields, such as software product business (Centers of Expertise, 2006). Also, one important local investment on urban technologies is the panOULU (public access network OULU) project, which provides in its coverage area wireless broadband Internet access to everybody (panOULU, 2007).

The software industry has grown rapidly in Finland, and the sector is one future possibility for economic growth. There are about 1,100 software companies in Finland employing 12,400 professionals and generating gross annual revenues of about 1.2 million euros. About one-third of the revenues come from international sales (Finnish Software Business Sector, 2006). In FUR Oulu there are about 14,000 jobs in the 569 ICT companies, and the sector provides 18% of all jobs in the region (SeutuNet, 2006). A total of 113 firms are classified as software companies, and they provide 1,200 jobs. Hence, in relation to the city scale, the amount of jobs in the ICT sector is high in Oulu. As the production process in the software industry is complex, product development requires a combination of several actors, also subcontractors and clients, know-how, and economic resources. First of all, the end products are developed in small series or single projects. Secondly, they are formed as a combination of several parts. Thirdly, they require a combination of many specialized fields of knowledge and know-how, which often requires intensive collaboration between specialized companies (Miettinen et al., 2006, p. 26).

COLLABORATIVE PRODUCT DEVELOPMENT: CASE STUDY OF OULU

Concepts, Issues, and Problems

The Technopolis concept differs from ordinary business centers by offering development services and creating a critical mass by gathering enterprises into the same premises. Technopolis in Oulu acts as a partner in regional development, where the main working method is forums. The idea of the forums is to strengthen the competitiveness and expertise of everyone participating in the activities by launching practical joint projects involving 20 to 30 companies. Ideally, these forums will result in new, successful products or practices (Launonen, 2006). Moreover, the cluster in Oulu is very compact physically and the traffic jams are very rare and slight, which makes temporary clustering easy. Almost any time of the day, for example, it takes only 20 minutes from airport to technology center (see Figure 1). Most software companies are located in Technopolis premises or the city center, so several meetings with different companies can be arranged during one day.

The previous studies (Ala-Rämi, 2007; Ala-Rämi & Inkinen, 2008) have shown that high-technology enterprises in northern Finland find ICT and face-to-face contacts both important and complementary to each other; also geographical proximity plays an important role in the enterprises’ inter-firm collaboration. This study aims to go deeper on these issues by using the case study, which provides glimpses of the experiences of globally successful software companies. They all have their product development in the city of Oulu and were born as spin-offs from another firm or project. The study examples are based on thematic interviews of managers of software companies (three firms out of six) that were chosen in the Global Launch program. The program leader was also interviewed. The idea of these interviews, besides to hermeneutic understanding of the issues studied, was also to find some ‘best practices’, and therefore the most advanced companies globally were chosen to interview. However, the study uses multiple sources of evidence and also previous study findings, including other authors (Männistö, 2002; Tervo, 2004; Simonen, 2007; Roberts, 2000), to support these results.

Global Launch is a tool of the Centers of Expertise Program, whose task is “to use in-
ternationally competitive knowledge and skills as a resource for business activities, creation of new jobs and regional development’ (Centers of Expertise, 2006). The program uses experience from a previous successful program (Niinikoski et al., 2006), but also takes into account challenges by including firms in very different development phases and orientations in business. So the Global Launch program was targeted toward ‘born global’ companies. The criteria in choosing these enterprises were that they already have a promising product, a large potential market area, and a competitive edge, which was good selection principle also for this study. Although the program did not start as planned, the interviewed companies participate in the larger ‘Global Cluster’ project in which the original program was merged. The aim of the interviews was to deepen some themes raised from a larger survey of software companies in the beginning of 2006 (see Ala-Rämi & Inkinen, 2008) which are only possible to study qualitatively.

All these companies are classified as small companies, having less than 50 full-time employees, and they were collaborating in product development. They have their main office with product development in Oulu, but they all have at least one sales office abroad, which serves international customers. The interviewed managers saw that internationalization is essential, since local markets are very small — only some 0.3% of software markets are in Finland. As the interviewed companies are already successful globally and have the potential to develop further, this study focus on ‘best practices’ already in use in these companies.

We understood in the very beginning to start producing software products for firms, mostly to software companies, but also to other firms, meaning test tools that have only one market area, which is the whole world. There are no local markets and never will be, one has to be global or there is no possibility of surviving. As a matter of fact, we got a half of our revenue from the U.S. already in the third year, and in the U.S. these games are usually make or break. We do have an American as our chief executive officer at the moment. We seek there that kind of feeling for business that is not found in Finland or at least is difficult to find...as we aim to be market leaders in our own trade, which means we must grow as big as tens of millions of euros annually.

— R&D Manager (author’s translation)

Social Networks of Software Companies

Along with a long history, the local technology center has had time to develop as an intensive growth cluster, in which the local culture is experienced as a supportive environment where ‘everybody knows everybody’ (Lappalainen, 2007; Tervo, 2004). So, there is still strong social networking in FUR Oulu, where information about new collaboration partners spreads mostly through informal contacts, such as former fellow students and friends (Männistö, 2002). Government arrangements of companies also play an important role in networking.

Well, mostly there has been some kind of existing contact, so [the parent company] is quite natural...In Helsinki there is one firm which is kind of owned by former [parent company] employees. Here in Oulu we have used one firm, in which there is such a situation that one of my old friends went to work there and this way that firm ended up as our subcontractor.

— Chief Executive Officer (author’s translation)

A strong community of practice has its drawbacks: there is a danger of lock-in and reduction of diversity (Oinas & Lagegendijk, 2005, p. 318), in case there is a limited number of skillful actors. Knowing the local actors and their know-how
Social Networking and Knowledge Transfer in Collaborative Product Development

is an advantage and creates flexibility, but it has some challenges. Collaboration with well-known people is fast and easy, even communication routines already exist. However, that is not always the most profitable solution in the long run.

A shared interest brings actors together. Around a broader discussion of cultural events or sports, it is easier to discuss more confidential issues and bring up new ideas. Many innovations are not that actively invented, but are found as a result of social interaction. Thus, temporary clusters act as platforms for new ideas (Maskell et al., 2006). The need is often brought up through knowledge of what other firms do and can do, not that much by searching for special know-how. Even though this kind of interaction takes place in informal situations, such networking can be supported by creating a temporary cluster in which many enterprises come from outside to meet enterprises in the region. Physically more distant relationships are often born at various meetings (Torre & Rallet, 2005, p. 53), exhibitions, and through other formal but personal ‘brought contact’. Web-based databanks are not seen as an effective form of networking when there is question of more equal and trust-based cooperation, since they cannot create togetherness. However, these kinds of solutions work better in finding new clients.

Communication Practices Between Collaborating Partners

There are some differences in practices and experiments depending on whether the company is small and product development is done more equally with another company or if the collaboration is a client-type solution. In both cases personal meetings are socially important, whether it is a question of mutual trust or transmitting the message that the company has time for an important client. In close physical proximity, meetings were common, even when dealing with issues that could easily be communicated with e-mail.

Face-to-face meetings create trust and understanding. Also, personal meetings are the best way to discuss topics in which there are many different phases (Torre & Rallet, 2005, p. 54), and some ideas can be shown with drawings or by “pressing the buttons on a keyboard.” These issues are not easy, if not impossible, to transfer by codified communication; the power to enunciate is also better in personal meetings. Time is a significant issue in product development; so many solutions are done on the least time-consuming basis. For example, even though Skype is often used, some find it too slow to use for communication, compared with a normal mobile phone.

Different virtual environments were not used that much, at least not so far (e.g., for conferencing), even though they can overcome some challenges of distant collaboration (Torre & Rallet, 2005, p. 54). They still have problems related to reliability, voice, and picture. The interviewed company managers said a much-used method is to use a phone or Skype and e-mail at the same time. One possibility is to do ‘multi-site’ work—using software that makes it possible to split the screen and a phone. However, that is not yet very widely used.

E-mail is absolutely the most important way, and we try to stress that very strongly. It has that saving grace, it has many saving graces, that it is filed, so we file e-mails, so then it is possible to go back and do searches...Many times it is possible to attach some clarifying material and such things. The only problem is that sometimes the message is, it's too slow to discuss with e-mail. After joining phone meetings with it, it is quite a good combination. Surely we do visit for clients, the closer the easier.

— R&D Manager (author's translation)

Even though the possibilities of Internet-based solutions have not replaced face-to-face meetings (Ala-Rämi, 2007), there are some things that are
preferable to manage with e-mail, such as most agreements. The main reason for that is documentation; there is evidence of what is agreed and what is done, but also some things like Web links are also easier to write than to say. However, e-mail is quite slow, especially between different time zones, at least in cases where there is a need for ‘discussion’. Also, the interviewees had some cultural differences, such as the North Americans mostly prefer telephone conversation instead of using e-mail. However, ICTs and face-to-face meetings are mostly seen as complementary to each other; different modes of communication are used in different kinds of situations and communication (Ala-Rämi, 2007; Gillespie, Richardson, & Cornford, 2001, p. 128).

...such as there are cultural differences in how Finns speak and how Americans speak, and how the talk is understood, for example when something is ready and what to expect in which time and so on...a democratic example is that it can be challenging to organize a meeting when the leader of the U.S. office is meeting with us every Sunday at 11 p.m. ...

—Chief Executive Officer (author’s translation)

Formation of Community of Practice and Geographical Proximity

Concerning the activities of a collaborative project, the first condition is to create a shared goal, idea, or concept (Wenger, 1998). Strengthening of the goal takes place in special circumstances requiring caring, trust, and commitment, but also freedom, vagueness, and even creative disorder. Trust is crucial for the emergence and success of a collaborative project (Miettinen et al., 2006, p. 56). Trust is developed best in unofficial situations and through a shared interest, which by choice is not business related (Oinas & Lagegendijk, 2005, p. 312).

If the case is that they have not come to talk with me about something particular, then there are right away certain matters that you can start to talk about, kind of outside the business...in which case there are certain ties that build trust. If there is plain business talk, the game rests on some CVs and really tough values...they do have [their] own place, but it is so much easier to call a guy with whom we have talked [to] four hours about what kind of movie it was and why, what was actually said there...than to a guy with whom we had a one-hour ‘transparency drill’, in which case there is nothing personal in the game. When there is something personal, the bond is strengthened.

—Chief Executive Officer (author’s translation)

Geographical proximity is meaningful in product development (Ala-Rämi, 2007). It is especially important in the beginning to be able to create an operational culture with a new collaborative partner. It is largely a question of trust, if there is no mutual understanding of what has really been agreed on and how responsibilities are divided. Once the CoP has been developed, knowledge sharing becomes smoother (Hislop, 2004, p. 40). Moreover, distance creates breaks in communication, and spreading of information is slower. Furthermore, cultural issues, language, and a time difference play important roles in distant collaboration. These issues are major problems created by distance between collaborative partners, not physical distance as such, but causing cultural and temporal distance. Cultural differences also have some influence on collaboration, especially in more intense product development where it is a question of creative work. However, client contacts were regarded as benefiting greatly from geographical proximity (see Oinas & Lageendijk, 2005), which was the reason all the interviewed companies had a sales office close to the main markets of their product, in Silicon Valley or London.
I haven’t heard that a client or someone would have said that it (being a Finnish company) would be a problem, but I would say that as we have an American CEO and an office there, it gives credibility... Well, there has to be some kind of representation or some kind of credible operations also in the main market area. But, it does not matter that the roots are in Oulu and even that the product development and such things are here, it is not a problem. It makes it sort of exotic also, and Finland is nevertheless considered a high-tech country, so no one thinks they cannot be good because they are from Finland, I have not met that.

— R&D Manager (author’s translation)

CONCLUSION

Due to its long history and experience in the ICT sector, FUR Oulu provides a good operating environment for software companies. The software sector is strongly based on a good ICT infrastructure, and the city of Oulu has put a strong effort in developing these services through the Technopolis concept, which has made knowledge diffusion between different actors smooth and easy. Besides different services offered by Technopolis, also the city of Oulu has developed wireless connections open to everyone around the region. Since the beginning of Technopolis of Oulu, there has been a strong community of practice, the ‘Oulu Phenomenon’, and it is still meaningful for high-technology enterprises in the region. These companies play an important role in regional development, and thus they are enhanced by a national program for regional development, such as the Centers of Expertise program. One of the core industries of the city of Oulu is software production. This study explored the use of face-to-face meetings and ICT in product development collaboration, and examined views on networking by using three case studies of globally successful software companies in Oulu.

The physical compactness of the city is an advantage; movement around FUR Oulu does not require much time. Furthermore, most software companies are located in Technopolis’ premises or the city center, so several meetings with different companies can be arranged during one day. Software companies in Oulu gain from strong local social networking, which is important in mutual collaborative work. It is also an effective information channel for new ideas and needed know-how. However, the number of skilled actors is limited, which is a problem for the interviewed companies. So, the small urban scale of Oulu has a two-fold impact.

Geographical distance has a meaningful impact on collaborative work, even though relative proximity is most essential in networking. Temporary clusters are essential for networking of actors. Oulu is well known in ICT, which facilitates access to global markets. However, the studied companies saw it essential to have a sales office in the main market area. Working at a distance in product development was easier to overcome, although it has an impact on practical arrangements, such as matching working hours with people in other time zones. However, there are more complex issues to overcome, such as creation of trust and mutual understanding. Cultural differences and different languages have an impact on communication and understanding. But it is also a question of different personalities and their ability to adjust in different situations.

Face-to-face meetings are needed throughout product development, but they are needed the most in the beginning of a project, especially with a new collaborating partner, to create shared goals and trust. ICT plays an essential role in the product development of software companies, not only as the means for routine communication, but as an important tool for transferring contracts and supporting documentation. As the number of local actors and markets is limited, it is necessary for a growth-oriented company to operate globally. Software companies in FUR Oulu have good
Social Networking and Knowledge Transfer in Collaborative Product Development

possibilities to be globally competitive in their field, but government support is seen as vital in entering the global market. However, there are also challenges in keeping globally successful companies as Finnish companies. Spatial proximity has been a supportive tool, and it can be put to better use by taking it into account in urban development.

FUTURE RESEARCH DIRECTIONS

Development towards more globalization is also a trend in Oulu. The city of Oulu has started a program with the aim of “making Oulu famous for its growth companies, which act globally and inspire the region of Oulu to be competitive and develop an innovative environment of different industries” (Oulu Inspiroi, 2006). This ‘Oulu Inspiroi’ (Oulu Inspires) program, including the above-mentioned ‘Global Clusters’ project, also provides possibilities for interesting research directions. What kinds of tools would be most valuable in supporting companies in networking and how does proximity support the development of the city? Moreover, how should these issues be supported in further urban development?

The ‘Oulu Phenomenon’ is the result of a long-term effort to support and develop the ICT sector. There are working national and local policies that have supported networking and clustering of the knowledge-based sector in the region. ‘Global Clusters’ is one of these practical tools. Time will show if the ‘Global Clusters’ program, as part of the CoP, will be successful. However, the interviewed companies argued that they have benefited from earlier global software programs. Still, there have been good inventions that have failed to make a breakthrough. In the opinion of the interviewed managers, it is a question of a lack of support in marketing. However, there are already tools for that in national and local policies, such as the Centers of Expertise program, which have been meaningful in successful clustering of software companies. So, these programs can be developed further to even better fit the needs of the companies in Oulu.

According to the interviewed company managers, one main challenge is to find more skilled professionals, which are few even in the whole of Finland. The first phase is to attract more and more delegates from international companies to come to the city of Oulu to make contacts. Technopolis has expanded its operation also to Helsinki, the capital of Finland, and more recently abroad. Just as the newest parts in FUR Oulu, they are also strategically located next to the airport or the city center. The existing urban technologies in FUR Oulu and ongoing development of technical infrastructures are supporting the knowledge city well, as long as their factual content is relevant.

ACKNOWLEDGMENT

I wish to acknowledge the financial support from the Academy of Finland research project 200771 and Tauno Tönning Foundation. I also thank two anonymous referees for their comments and suggestions on ways to improve the chapter.

REFERENCES


Social Networking and Knowledge Transfer in Collaborative Product Development


Social Networking and Knowledge Transfer in Collaborative Product Development


ADDITIONAL READING


interaction and local development. Aldershot: Ashgate.


*This work was previously published in Creative Urban Regions: Harnessing Urban Technologies to Support Knowledge City Initiatives, edited by T. Yigitcanlar, K. Velibeyoglu, and S. Baum pp. 65-79, copyright 2008 by Information Science Publishing (an imprint of IGI Global)*
Chapter 6.5
Reframing Information System Design as Learning Across Communities of Practice

Kevin Gallagher
Northern Kentucky University, USA

Robert M. Mason
University of Washington, USA

ABSTRACT

This article frames the requirements definition phase of systems design as a problem of knowledge transfer and learning between two communities of practice: IS designers and system users. The theoretical basis for the proposed approach is Wenger’s (1998) framework for social learning, which involves three dimensions: alignment, imagination, and engagement. The article treats the requirements definition task in systems design as a set of activities involving mutual learning and knowledge transfer between two communities of practice (CoP) along these three dimensions. In taking this approach, the article maps the results of past research on the systems design process onto this CoP framework and illustrates that the proposed framework encompasses the same activities used by traditional methods of requirements definition. However, this approach focuses attention on the learning that must take place between the two CoPs and thereby helps resolve some of the inherent shortcomings of prior efforts and approaches. The framework provides both a more encompassing conceptual lens for research on improving the requirements definition task and practical guidance for managers who are charged with a systems design project.

INTRODUCTION

Requirements definition is a critical step in systems development that requires the identification
of information needs and knowledge of a system’s processes (Nelson & Cooprider, 1996; Vessey, 1994). Historically, researchers examined the requirements-definition stage of system design as a process of inquiry (Boland, 1978; Salaway, 1987). Problems with identification, articulation, and communication of information needs have long been identified with the challenges of information system design (Boland, 1987). There have been different approaches in attempting to meet these challenges, but none has completely resolved the issues.

Land (1998) notes that because systems are so different, a contingency approach—using different methods for different types of systems—is appropriate. Others have suggested more structured analyses of the design process itself, establishing metrics for requirements engineering (Costello & Liu, 1995) and developing tools for each aspect of the problem (Nature_Team, 1996). Some researchers have suggested that the process of design must remain flexible and that a management structure that encourages an evolutionary design process is associated with greater effectiveness (Ravichandran & Rai, 2000). In considering software project risk and software quality, organizational issues as well as technical issues are important (Wallace, Keil & Rai, 2004). Others also emphasize the critical nature of human-intensive dimensions of the process (Tamai, 1993). It also has been noted that evolutionary designs are necessary as complexity increases (Mens & Eden, 2005). Larman (2004) argues that an “agile” and iterative design process is key to software development success.

The approach that we want to explore in this article emphasizes these human-intensive dimensions of the design process. Although the design process involves many actors (Lamb, 2003), we want to focus on two roles: the designer (who has technical knowledge) and the user (who has knowledge of the application and context of use). The conceptual approach is one that considers requirements definition as an instance of knowledge acquisition (Byrd, Cossick & Zmud, 1992).

Recently, organizations and researchers have begun investigating the potential of knowledge transfer (Alavi & Leidner, 2001) to make organizations more effective when engaging in information intensive work. Such knowledge transfer is necessary because clients are not sure what is possible and are unclear about their needs, and IT designers thus are unable to work toward an outcome that meets clear specifications (as in designing a product for production) (Larman, 2004).

To date, however, conceptualizations of knowledge transfer in software development do not completely capture the complexity and richness of this process by which clients and designers work together. As Polyani (1966, p. 4) says, “We know more than we can tell.” Regardless of how well we articulate knowledge, it always contains a tacit dimension. Hence, simple inquiry is insufficient for the requirements definition process because it is able to access only explicit, leaky knowledge (Von Hipple, 1994). The information transferred through traditional elicitation approaches is only part of what someone knows, and it rarely includes how or why they know it (Lanza & Mathiassen, 1985).

Because of the tacit dimension of knowledge involved in most tasks and processes, it is difficult, if not impossible, for people to articulate exactly what it is that they need prior to design. Even if they can articulate what they need, the system development effort is hampered if the system developers do not understand why and how users need what they need. With an understanding of the why’s and how’s of information, developers can be more innovative in their delivery of requirements. For example, unless developers understand which information is used together and how it is connected, they will be unlikely to find ways to combine and simplify tasks.

If we keep the traditional concept of inquiry as the basis for eliciting requirements, the effort must
always be incomplete. The language or metaphor of “inquiry” or “capturing requirements” is part of the problem. The choice of metaphor can inhibit the range of approaches to information systems (Mason, 1991), and recently the metaphor of “engineering” has been questioned as the basis for approaching software development (Bryant, 2000).

What is needed is an approach that engages users in such a way that they can elucidate not only what they know, but also how and why they know it. In addition, users must learn about how systems in general, and this system in particular, will be developed. An improvement over current approaches is to engage designers so that they really understand what the users are telling them, and to ensure that they effectively share their knowledge of both the systems and the processes for which they are developing technology (Boland, 1978). To accomplish this, both users and developers must share their knowledge and learn from each other (Boland, 1978; Churchman & Schainblatt, 1966). In short, what is needed is a process of mutual learning in which designers and users engage in a process of learning from each other.

In order to work toward such a process, we turn to the concept of a community of practice (Lave & Wenger, 1991). The concept, first used to help understand situated learning, refers to a group of people defined by an interest in a specific subject or problem who collaborate to share ideas and resolve issues common to the group. In this article, we posit that information system (IS) designers and IS users belong to two distinct communities of practice (CoPs) as they work together in a systems development context. The system designers comprise one community and the users for which the system is being designed comprise the other (Churchman & Schainblatt, 1965). The term “users” may cover many roles, including managerial, and in the following we may use “managers” to provide a specific context for the discussion or simply say “users” for the generic situation. The requirements definition phase of IS design requires that these two groups engage in mutual learning both to share knowledge and to develop new mutual understandings of the possibilities as a system design emerges. The knowing of managers (or other users) is formed by a very different practice than the knowing of IT professionals (Brown & Duguid, 2001), resulting in different ways of knowing (Boland & Tenkasi, 1995). Knowing is not easily shared across these two groups because of their different histories and perspectives.

The article extends Wenger’s (1998) design for social learning within a CoP (p. 236) to learning across two communities of practice. Wenger’s model comprises three dimensions: alignment, engagement, and imagination. We apply this framework of social learning across the two CoPs (i.e., managers and designers) and show that the framework not only is consistent with prior research on the information system design (ISD) process but also suggestive of how the design process can become more effective.

The remainder of the article comprises three sections. The next section, Social Learning, reviews the concept of social learning and distinguishes it from simple knowledge transfer. This section further illustrates why the characteristics of social learning are the characteristics required for an effective requirements definition process.

The subsequent section presents an architecture for social learning. It reviews the three dimensions of Wenger’s model for social learning in a community of practice and notes that factors considered essential to the requirements definition phase of the system development process can be mapped onto these dimensions.

The final section summarizes the application of the model for information system design, outlines issues in implementing the architecture, and discusses the implications for research and practice.
SOCIAL LEARNING

Knowledge and Knowing

Viewed from a social learning perspective, knowledge is social, not individual (Brown, 2002), and is a matter of our competence with respect to some valued enterprise (Wenger, 1998). Knowledge encompasses aspects of experience, action, and the accomplishment of some activity. In this perspective, knowledge is more than stored information, even more than codified and structured information.

It is helpful to distinguish “knowing” from “knowledge.” In common usage, “knowledge” brings with it the context of an accumulation of facts and “know what.” On the other hand, use of the term “knowing” conveys more of a sense of engagement in a process of learning and becoming. Knowing is not only an accumulation of facts, but an accumulation and assimilation of experiences and learning that alter who we are as well as what we know. As we learn about, we also learn to be (Brown & Duguid, 2001). Levels of participation or nonparticipation reveal a sense of relation, familiarity, and belonging to the situation. These aspects of identity development affect one’s modes of involvement and willingness to engage in activities. The attributes of identity are developed, and reciprocally supported, through the process of learning.

We are constantly engaged in accomplishing activities and we typically become more proficient with experience. Our knowing, therefore, resides in our practice (Cook & Brown, 1999), which is how we accumulate experience and situated ways of knowing. Participation is our “knowing in action,” so a design process can benefit from participation by those who know the existing process.

Apprenticeship and Communities of Practice

Communities of practice provide a theoretical framework for understanding how the dynamics of social structures influence learning and practice. The origins of communities of practice are grounded in apprenticeship and their ability to support situated learning through social coparticipation (Hanks, 1991; Wenger, 1998). Understanding how apprenticeship promotes learning within a community can help in understanding how learning can occur across communities. Learning through practice is exemplified in apprenticeship, whereby a community regenerates itself through a process of participation and training. When the transfer of explicit knowledge is supported by sharing the practices in which it makes sense, it is possible to understand its meaning in a way that makes it most useful, actionable, and purposeful.

Participation is the key to developing and sustaining communities of practice, but observation also plays a critical role by allowing one to participate on the periphery of a community. While the master remains engaged in the craft, the apprentice learns by means of “legitimate peripheral participation” (Lave & Wenger, 1991), observing the practices and developing the tacit dimension of knowing. By engaging in practice, the master is able to communicate more than what can be said and the apprentice is able to learn more than what can be heard.

The roles of a community’s members are not rigidly defined; however, a key to apprenticeship is that members can engage in varying roles simultaneously. Legitimate peripheral participation supports learning without requiring members to acquire all of the necessary knowledge for full membership in that community. In this way, a member can span the boundary of a community and still be able to observe, understand, and learn.
Participation and observation (nonparticipation) together define our levels of engagement as we employ these modes of engagement in attaining knowledge of our own communities and other communities from which we may learn.

To participate on the periphery allows one access to a community through passive modes of observation and similar activities that do not overly dominate or eliminate participation altogether. This is exemplified by active listening. Conversely, marginality is a mode of nonparticipation that stifles participation, as exemplified by listening in on a conversation in which the terminology is foreign. In both cases, there is exposure to a community of new information. The first situation offers a mode of nonparticipation that facilitates learning. In the second, the learning becomes stifled.

Apprenticeship is offered here as the starting point of a model to use when designing a way for two communities to come together, share knowledge, and learn. In the case of requirements definition, it may be enlightening to consider the IT professionals to be apprentices to the user community (Beyer & Goltzblatt, 1995) and the users to be apprentices to the IT community. Through apprenticeship, the participants learn about practice. To make “know that” useful requires appropriate “know how” that we can only gain through practice in both senses of the word: to do and to improve (Brown & Duguid, 2001). The need to involve users in system design has been a long held belief for requirements definition; however, involving IT in the practices of the users is less established, but has the potential to be quite revealing.

**Knowing as a Challenge for IS Design**

The challenge for IS design is to enable participation by both users and designers in a process that both conveys knowledge and enables knowing. In any social process, our experience informs our knowing, but the structure of the relevant social arrangement orients and guides our experiences. We develop personal histories from experience, and, through practice, we negotiate and form the meanings we give to things. Identity is how we see our role in the process that then influences our level of engagement.

Both users and designers must be engaged in the process of knowing about the other’s needs and the opportunities afforded by alternative designs. Engagement is a social process, one not naturally emerging nor easily brought about by fiat.

Prior researchers have acknowledged the social aspects of information system design. For example, political arrangements in organizations and dynamics of relative power among design participants can influence the degree to which they feel safe participating in activities (Hayes & Walsham, 2001). Others have noted the importance of the relationship between users and designers in the design process (Beath & Orlikowski, 1994; Hirschheim & Klein, 1994; Ives & Olsen, 1984). Mutual learning can only take place when users and designers are able to engage in effective modes of participation when they are involved in requirements definition activities. Through appropriate modes of participation, requirements are informed both by what the participants know and by how and why they know it.

What we want to accomplish in the IS design process is to enable both designers and users to engage in a learning process that enables safe but effective participation in the design activities. The apprenticeship model of social learning in a community of practice, coupled with Wenger’s (1998) architecture for social learning (p. 237), provides a model for such a process.
AN ARCHITECTURE FOR SOCIAL LEARNING

Processes of social engagement differ in their ability to support learning. Figure 1 illustrates Wenger’s (1998) architecture for social learning in a community of practice, showing arrangements for alignment, imagination, and engagement. When engaged in practices across communities, the support of such social arrangements can facilitate a sense of belonging for members that helps to ensure participation and learning. Collectively, the three dimensions enable members of the community to learn with each other. The architecture provides a framework for considering the requirements of the process of IS design, particularly the requirements definition phase of this process.

The three dimensions are interdependent and, in fact, have inherent tensions among them. For example, a structure that assures close coordination for alignment can inhibit imagination and engagement. Conversely, imagination without alignment can result in activities that do not contribute to the overall learning process.

Alignment ensures coordination of the community’s activities and resources toward achieving an objective. In the case of system design, this means achieving a fit between the local needs of a business unit and the new design. Learning is important to such a practice, as working through the development of a design raises new questions from which participants can learn about the existing system (Kyng & Morten, 1995). However, there are many ways to establish a fit during design, some of which may inhibit future flexibility (Allen & Boynton, 1991) and lead to rapid obsolescence or an inability to adapt. Alignment alone is not enough to support learning. A community of practice, and perhaps more importantly, the interaction between two communities, must also support what Wenger labels imagination.

In Wenger’s terms, imagination provides the opportunity to create new possibilities. The ben-
Reframing Information System Design as Learning Across Communities of Practice

Advantages of creativity and innovation in the development of systems are well established in the field of information systems (Cougar, 1996). However, creativity without the appropriate information as to the possibilities and constraints of a design or an appropriate vision of the future may lead to poor implementation (Zmud & Apple, 1992). Therefore, a process of mutual adaptation is necessary to achieve an improved design and a successful implementation (Leonard-Barton, 1993). Mutual adaptation can be viewed as an accommodation of both alignment and imagination that can be accomplished through learning.

The third dimension of Wenger’s design, the concept of engagement, represents active involvement in the negotiation of meaning during the learning process. Engagement requires appropriate modes of participation. Appropriate modes of engagement by participants can ensure a balance between alignment and imagination. By supporting participants’ engagement in a social process, it becomes far more difficult for any one set of perspectives or goals to dominate the process. However, engagement is difficult to ensure. Psychologically, individuals must feel it is safe for them to engage in an activity, that it will be a meaningful investment, and that the personal and professional resources that are necessary for participation are available to them (Kahn, 1990).

In the apprenticeship model of learning, engagement in practice by the bearers of knowledge will assist in their grasping, articulating, and transferring knowledge to others. Furthermore, peripheral participation by those who are in apprenticeship roles facilitates their ability to appropriate the meanings conveyed in specifications by those engaging in the design process. In sum, engaging in the practice of design can help us to understand and transfer what we know.

The challenge in employing Wenger’s design for learning in the IS design process is balancing the tensions inherent in each of the three dimensions (See Figure 2 below.). Traditionally, IT departments have been concerned with alignment. They feel a need to align themselves with business strategy, with user functions, with business processes, and with organizational objectives (Chan, Huff, Barclay & Copeland, 1997; Reich & Benbassat, 1996). They have perceived standardization of

Figure 2. Wenger’s architecture for social learning adapted for requirements determination
systems and control over the design process as the key to alignment (Kirsch, 1996). However, with excessive alignment, imagination is curtailed and success becomes elusive (Baskerville & Stage, 1996). Excessive imagination and creativity, on the other hand, can lead to idiosyncratic designs and unworkable heterogeneity, reducing the ability to share, diffuse innovations, and efficiently develop new systems.

The quest for alignment makes it necessary for IT and users to work together. Yet, mutual engagement in these activities is in direct tension with the development of functional expertise within the organization. It also raises organizational tensions around leadership and decision making, between working together and understanding who has jurisdiction. As many IT scholars have claimed, partnership is key to IS success (Henderson, 1990). Partnership addresses the tension between conflicting power bases in the organization, but a successful design environment must also address the tension between imagination and alignment. Wenger’s design for learning recognizes the need to balance these tensions in the design of a learning environment that will support the appropriate modes of engagement and thereby facilitate and balance the tension between alignment and imagination.

The information system design (ISD) literature is consistent with this three-dimensional framework of learning in a community of practice. The literature on use of methods and controls is consistent with Wenger’s dimension of alignment. The need for imagination during system design is supported by the ISD research on learning by doing, performed through modeling and prototyping. Research on user participation and involvement supports the need for engagement.

The next section discusses the commonalities between Wenger’s framework and the ISD literature, explains further the facilities used to implement the model, and highlights the challenges of applying the principles and balancing the tensions in the ISD setting.

**Alignment**

In Wenger’s model (Figure 1), alignment is comprised of three facilities: convergence provides a common focus; coordination provides methods and procedure; and jurisdiction provides policies and procedures. Similar themes are found in the IS literature, which describes the need to pursue objectives, employ standard methods, and control the design process.

Objectives are defined based on user needs and the possibilities the system affords, negotiated through a social process of mutual adaptation (Leonard-Barton, 1993). Through methods, activities are organized to achieve these objectives. Yet, methods often give primacy to designers over users (Beath & Orlikowski, 1994) or to problem solving over alternative rationalities (Boland, 1979). Thus, formalizing methods serves to structure activities, thereby imposing controls that promote and suppress certain behaviors.

The IS design process requires more than obedience to method, it must also be a reflective activity (Iivari, 1987). System design requires iterative and adaptive planning as the acquired experience and learning from each sub phase informs the ongoing effort (Iivari, 1987). Thus, development efforts must recognize the need to balance iterative development with complimentary methods of control.

Control is complicated by the fact that coordination and communication necessary to complete the tasks spans organizational departments (Kirsch, 1997). A complex relationship exists between efficiency and quality. Different strategies of control lend themselves to different situations. Informal methods of control are used when understanding of the process and outcome measures are inexact (Ouchi, 1979) and depend on social strategies (Eisenhardt, 1985). Self-control relies on the motivation of individuals, and in IS it is strongest among personnel who take part in defining work procedures (Kirsch & Cummings, 1996).
Clan is another informal control characterizing a group that is dependent on one another to achieve a common objective (Ouchi, 1980). Boland (1979) distinguishes “control over” from “control with” in order to distinguish bureaucratic processes from those where shared values and beliefs serve as the basis for coordinated action. Through appropriate selection of members and use of appropriate training and socialization processes, individuals in a group can develop a stronger sense of identity and commitment (Kirsch, 1996). Appropriate activities serve as rituals to reinforce acceptable behaviors. The development of identity and commitment are equally vital components in apprenticeship, where methods of training and socialization are employed to facilitate learning directed toward achieving objectives. In the end, certain modes of control, especially in excess, can have a negative impact on both engagement and imagination.

**Imagination**

Imagination is comprised of three facilities (Wenger, 1998): orientation provides a location relative to the organization; reflection provides facilities for comparison and evaluation; and exploration provides opportunities for envisioning, creating and testing new ideas. Again, similar themes are found in the IS literature that emphasize a need to understand situational context and explore possibilities through processes that support discovery and innovation.

Systems requirements are the first step in defining possibilities for new information technologies. As previously noted, the development of system requirements is often conducted through inquiry and other elicitation methods, intended to identify user needs. Yet, this step is highly prone to unintended influence through bias and error (Salaway, 1987). In addition, interviewing can close down inquiry, if motivated to solve problems rather than explore the appropriateness of different possibilities for an organizational situation (Boland, 1979).

Another method of elicitation is the use of design tools, such as computer assisted software engineering (CASE) or use cases, which help users describe the organizational situations in which they are involved and convey what they know (Lanzara & Mathiassen, 1985). Such tools help users move from the current situation to the enactment of a new one (Lanzara & Mathiassen, 1985). For instance, redirecting the attention away from the existing system and refocusing on a new realm of possibilities (Wastell, 1999) is one such example. To reframe a situation leads to a discovery of possibilities.

Drawing on theories of action, Lanzara and Mathiassen (1985) are critical of most existing tools. They suggest that better tools are needed to bring the “background into the foreground” by offering ways to make procedure more explicit, thereby relating descriptions to the actual settings. This advances a user’s opportunity to question the appropriateness of existing solutions, challenge their adequacy, and take more responsibility for initiating a more positive action.

Similarly, prototyping enables a common reference point for users and developer, which helps to develop a mutual understanding and draw out and refine user requirements (Alavi, 1984). Reifying what is known helps to reveal what is not fully understood and offers alternative active techniques for articulating and defining existing situations that allow the designs and users to engage in a social process of simultaneous definition, design, and discovery.

To summarize, there is great support in the IS literature for imagination in the design process. There also is great concern regarding the mutual understanding of context among users and designers, which is necessary to define the present situation and to inform an effective process of discovering new possibilities. Furthermore, the tools we have available are limited in their ability...
Engagement

Engagement is facilitated by *mutuality* though: reciprocity and interdependence, which provides opportunities to form and share perspectives; *competence*, which supports the occasion for applying knowledge and exercising judgment; and *continuity*, which supports mechanisms of memory, which work as repositories for information created through social encounters. Similarly, IS research has found that the design process and subsequent adoption of a technology benefit from user involvement, which enhances confidence and acceptance of the technology (Hartwick & Barki, 1994; Ives & Olsen, 1984).

Including users in the design process involves more than their presence, however; it also requires overcoming conflicts of interest and effects of authority (Mumford, 1983, p. 31). Often users are given relatively passive roles, despite being required to sign off on requirements (Beath & Orlikowski, 1994). As a result, patterns of user participation vary according to who controls the selection of features and the coordination of activities (Kirsch & Beath, 1996). Participation, therefore, is often token or compliant, rather than shared.

The participative design (PD) movement highlights empowerment of users as important in their ability to influence the overall design process (Hirschheim & Klien, 1994). Joint application design (JAD) is designed to actively engage users and designers in activities, often away from their usual work settings and organizational roles, recognizing the importance that altering structures has on the dynamics of a design process.

One way that JAD involves users is through modeling and prototyping. Models support the development of shared understanding between users and developers, thereby building confidence in what is established as knowledge during inquiry. Prototypes serve to cultivate user participation and help build credibility and establish good working relations.

As with alignment and imagination, there is great support in the IS literature for engagement as a key component of RD success, but there is concern about our ability to successfully nurture it. The following section discusses the results from several researchers who have provided insight into approaches and tools that may help to overcome some of the most frequently encountered impediments to successful alignment, imagination, and engagement in the RD and ISD processes.

Summary: Designing an ISD Process for Social Learning

Table 1 illustrates the parallels between Wenger’s model of social learning, an apprenticeship model, and the factors judged important to successful design from the ISD literature. What remains is the construction of social learning processes that enable a project to implement and balance these factors.

Designing Structures for Social Learning

The social process of ISD, whereby users and developers collectively define requirements, is complicated in that participants may differ in their backgrounds, cognitive styles, personality, job characteristics, organizational circumstances, affiliations, or location (Robey & Markus, 1984). As one might expect, this leads to differing interpretation, confrontation, and misunderstanding.

A major concern, therefore, is that the discourse required for an effective process of inquiry is distorted in traditional organizational structures. Altering the structure of a design process through choices regarding the number of participants, the activities conducted, and the role of users and
designers can offer greater opportunity for users to raise issues, question the appropriateness of choices, make and demand commitments, and express attitudes and concerns (Hirschheim & Klien, 1994).

Nonaka and Takeuchi (1995) show that through organizational design actions enacted through structured mechanisms, knowledge creation is facilitated. Churchman and Schainblatt (1965) note that researchers and managers viewed problem-solving differently and propose a dialectic approach to manager-researcher communication. Boland (1979) found that structuring problem-solving processes in ways that are enabled by differing rationales encouraged participants to share their knowledge and arrive at different types of solutions. Different processes for understanding and applying knowledge are influenced by the controls placed over the process of user-designer interaction.

Altering the structures of user-designer interaction may provide benefits beyond improved specifications or overcoming user resistance. Bringing users into close interaction with designers promotes social sense making (Boland, 1984) and shared understanding (Nelson & Cooprider, 1996). Information analysis, extended to include the context within which data are interpreted, also provides design mechanisms that can support self-reflection (Boland, 1979).

In summary, the IS literature shows support for the notions resident in Wenger’s design for social learning. However, it does not explicitly address the tension inherent in alignment, imagination, and engagement. By exploring structures that balance these tensions, the process of system design may benefit, as we demonstrate in the next section.

### Resolving and Managing Tensions as a way of Explaining Successes and Failures

In this section, we revisit three previously published studies to illustrate the value that the social learning model adds to the literature in information system design. The studies we examine used three different methods: controlled experiments,
field research, and comparative analysis of previously collected case studies. The authors of each study include in their data and their analysis consideration to all three of the dimensions in our model: alignment, engagement, or imagination. However, in each study, one of the dimensions is used as the primary focus of investigation and therefore becomes a primary explanation for the results and the authors’ understanding of what distinguishes effective from ineffective design processes.

For each study, we offer an alternative explanation that encompasses the original differentiating characteristics but offers additional insights into how to manage the process of information systems development. With these insights, the tensions among all three dimensions are included in the management framework rather than emphasizing one dimension of the model over the others. Consequently, this research contributes to the research literature by emphasizing the need to balance the dimensions in the model, and the result is a broader perspective on the practice of managing the system development process.

Boland (1978)

In the first study we examined, Boland (1978) addressed the question of how a user is involved in system design. Employing an experimental research design, “two radically different processes of interaction between a systems designer and a manager were compared in an information system design exercise.” The author refers to the first as traditional and the later as an alternative. The first process was conducted by having the designer interview the manager. Afterward, the designer made suggestions for the design. In the second process, there was an initial sharing of information followed by a period of mutual suggestion and critique by each participant.

In the study, the structure (methodology and control) influences the interaction between the user and the manager. Based on how this protocol or interaction takes place, the problem solution is derived. The author states “the alternative interaction approach produced higher quality designs with important implementation advantages.” From these results, Boland suggests that the protocols “may help to define different problems, and thereby produce different, but equally rational, solutions.” As stated in the article, “the structure of their interaction is defined as the protocols or accepted patterns they follow in punctuating their exchange of messages.” Using the model proposed in this research, the causal direction between the dimensions inferred in this statement is quite clear. Alignment, which includes methodology and control, sets the stage for how users and designers will engage in the process. The author also states, “the structure of their interactions serves as the context for the generation, interaction, and interpretation of ideas between them. The protocols they followed are seen as a source of rationality in the design process, as they guide in bounding the problem space, drawing inferences, and defining an ‘appropriate’ information space.”

From the social learning model perspective, the protocol influences design activities, or, in other words, the causal direction inferred is (the words of our model) “alignment affects imagination.” A richer explanation, employing the proposed social learning model, is that the alternative process resolved tensions in the process and, compared with the traditional process, struck a better balance in what was emphasized as important. Protocols, as a part of a prescribed methodology, served to organize the process or inquiry, but as exhibited in the study, they may not engage users in ways that promote knowledge transfer or afford the ability to imagine the design of an innovative system. The alternative process certainly undertook a different rationality, as the author went on to discuss, but it also undertook a more balanced approach across the three dimensions of the proposed model. By having participants...
first share information, then describe solutions and discuss pros and cons, both engagement and imagination were enhanced.

Wastell (1999)

In the second study we analyzed, Wastell (1999) also viewed social structure as a vehicle for learning and as a critical element in information system design. The study involved field research and employed an action research approach. Opportunities for design to influence the structure of interactions between participants or the development environment and the way it is managed and controlled are surfaced in the study. Three case studies are described. In each case, the author suggests approaches, and then describes how these changes improved performance. Using a theoretical perspective from psychology, the author proposed that antilearning defenses create difficulties for information systems design.

Stress and anxiety, arising from the group defenses of those involved in the system development effort are at the center of this approach. For example, following traditional protocols allows participants to adhere to prescribed behavior rather than engaging the participants from other groups in a common set of tasks. The approach to improving method and involvement are to reduce stress and defense avoidance behaviors. As prescribed by the author, modeling and other activities of design provide the design process with transitional objects, which facilitate involvement by designers and managers. As a result, the use of these objects promotes a learning environment, or transitional space. This space is a structure that promotes learning, allowing designer and managers to safely engage in activities. Once created, management must try to support the existing environment, which is quite different from the idea that methods and controls are needed to ensure the effectiveness of a process.

The author sums up the problem in this case as a situation in which “a highly prescriptive methodology is imposed by fiat” compared with an alternative that works better. The alternative, described in a similar case study, is a process in which practitioners create their “own methodological ideas before consolidating these experiences into new working practices.”
Using the social learning framework model, we might explain the findings as one in which there was a shift in emphasis from a strong adherence to an ineffective method toward greater emphasis on both modeling and engagement of users and designers. The emphasis has moved toward greater engagement of users and what is described as a more “experimental” approach to development, and created a transitional space for those involved in the project. In this social learning explanation, the shift is toward a greater balance between the engagement and imagination dimensions of the model.

Kirsch and Beath (1996)

In this third study, the researchers again analyzed participation in system development projects. Kirsch and Beath (1996) analyzed data from eight case studies as a way to understand the relationships between different modes of involvement, the processes of coordination between designers and manager, and the product of the design effort. The process of user involvement is examined in terms of how involvement is actually enacted, as opposed to how methodologies may be prescribed in the literature. The authors examine this relationship from several directions in discussions at the end of the article, but their initial focus in terms of theoretical direction of causation is that participants will enact a methodology, rather than just following it blindly. The product of design is also examined by looking at who makes the selection of system features.

These authors focus on the effect of involvement, or engagement as defined in the social learning model. The study identified three different patterns of user participation: token, shared, and compliant. Each is shown to have different implications for a set of outcomes: task-system fit of the resulting system and management of conflict among these outcomes.

The authors propose that these patterns (token, shared, and compliant user involvement) vary in terms of who contributes technical and domain knowledge to the project and who controls feature selection. Involvement by users also influences the coordination mechanisms employed, the level of conflict, and how these conflicts are resolved. The data are examined in a number of ways, but the emphasis is placed on understanding how engagement influences the process by which the project is coordinated (alignment) and by which the system is designed (imagination).

To the first point, the authors state that the “enactment [of user involvement] is as much the choice of the client as it is the choice of systems developers.” This leads the authors to suggest that methods might focus less on the way that designers should try to engage users but more on developing mechanisms to elicit user knowledge. To the second point, regarding the influence of user involvement on the modeling and design process or the effect of engagement on imagination, the authors conclude that high user involvement is not always necessary for task-system fit. The findings show, however, that when the system design features were derived by a process of negotiation between users and designers, the process yielded two of the three highest quality designs in terms of task-system fit.

Using our social learning model, an alternative explanation is that the processes that engaged users and designers took into account what participants do (imagination) and how the learning process is managed (alignment). As a consequence, the engagement process struck a better balance among the three dimensions in the proposed model. Engaging in a complex and often ambiguous process can be troublesome, especially if there is a great deal of conflict. In this study of eight cases, the three cases that had the highest rating overall had the least conflict. And while selection of features plays an important role in
determining fit, as pointed out by the authors, feature selection does not have to be controlled by users. But their engagement and participation in negotiating the design process yielded the two cases with the highest ranking overall.

**Summary: Additional Insights Using the Social Learning Model**

As noted, each of these studies discussed the interaction of the three dimensions in the proposed social learning model. In each study, causal inference was not the primary object of their investigation, as each study undertook a process perspective in their research approach, which we acknowledge and appreciate. These studies collected and reported rich sets of data and offered thoughtful and insightful analysis, enabling us to revisit them from the perspective of the tensions inherent in the social learning model.

By using these particular studies as a baseline, we are able to explore alternative explanations for the results of each. That is, we can offer different, richer explanations for how the studies unfolded, and more importantly, how the results came about. The studies started with different foci: Boland with control and method, Wastell with modeling and design, and Kirsch and Beath with user involvement. As a result, the three studies examined their data using different theoretical perspectives and thereby derived very different causal explanations of what occurred.

The social learning model does not resolve the conflicting direction of causation implied by these earlier studies. Instead, the learning model provides a “both/and” rather than an “either/or” explanation. Our model shows the relationships between the dimensions in the model as interrelated, with tensions that need to be resolved, rather than the need to optimize any one of the dimensions.

Finally, we note that all three studies demonstrate the need to balance engagement and imagination. While Boland found that more engagement was related to more and better designs, and Wastell found that the activities of design helped to engage the two communities in a more effective social process, Kirsch and Beath found that varying combinations of user and designer involvement has differing effects, dependent on many other factors. As a consequence, simply trying to increase involvement of users does not necessarily lead to better designs. Alternatively, the availability of an innovative or a structured design processes will not insure users’ participation, nor can the design process insure that those users will acquire a sense of involvement.

**DISCUSSION AND CONCLUSION**

This article has posited that the necessary exchange of knowledge between the designers and the users of information systems during the design process may be viewed as exchanges between two communities of practice. The article further suggests that the apprenticeship model of learning is an appropriate model for this knowledge exchange in which there is mutual learning. In this model, the user is the apprentice to the designer (when learning of technical capabilities) and the designer is the apprentice to the user (when learning of the system requirements and use). Apprenticeship conceptualizes knowledge as constituted in action, so separating the tacit from the explicit is not possible. Our knowledge exists in our competencies to perform or engage in action, and these competencies are formed in our learning through the negotiation of meaning. It is through this mutual apprenticeship process that we can grasp and articulate the meanings of the information we need to transfer. Mutual apprenticeship is a means by which both communities can learn, enabling the sharing of both the tacit and explicit dimensions of each knowledge domain.
Apprenticeship is the most accessible example of the benefits of both participation and nonparticipation and how the evolution of shared histories and development of the roles played by the apprentice and the master bring about the transfer of knowledge. If a master is asked to merely explain certain knowledge to the apprentice, the conversation would be flawed by the misunderstanding arising from the differing interpretations of the words that were used. However, through a process of engagement in a situated practice, development of competencies, and the evolution of identities, knowledge is transferred across communities via the apprenticeship model. The architecture for learning provides focus for balancing the tensions among alignment, imagination, and engagement.

Changes in structure can allow a community to move beyond just transferring “know what” to sharing greater degrees of “know how” and “know why.” By engaging in common practices, communities can understand greater degrees of the tacit dimension that exists in all knowledge (Brown & Duguid, 2001). However, even communities who work at the same practices can have differing ways of knowing. These differences result in knowledge being sticky and difficult to transfer, and they demonstrate the social dimensions of knowledge arising from the situated ways in which practices may develop.

By altering modes of participation in line with an apprenticeship model of learning and by supporting alignment and imagination through each community’s engagement in practice, managers can reduce the difficulties of sharing knowledge across the communities of users and designers. Perhaps of greater value, the knowledge shared across these communities, through altering modes of participation and balancing dimensions of the learning process as defined in Wenger’s model, can result in greater degrees of “know how” as well as “know what.” Furthermore, communication of related practice associated with the knowledge transferred and communication of the historical reasons for the practices used and the meaning appropriated in activities express aspects of a “know why” dimension embedded in the community’s way of knowing. Each of these changes should yield more creative designs and systems that have a greater degree of alignment with the organization’s goals.

We examined several cases of success and failure from the perspective of the causal explanations put forth by the investigators and using the model proposed here. In particular, we found it enlightening to compare the balance among the three dimensions of the model in successful and unsuccessful projects, which showed how the former had exercised a better balance than the latter. Ultimately, alignment, imagination, and engagement must all be supported, but focusing too heavily on one dimension may disrupt the balance among them. Similarly, the various modes of participation in the apprenticeship model have the potential to benefit a learning process, but their implementation as changes to an overall process may have very subtle influences based on the number of participants and their role in the organization.

The model provides important insights into how apprenticeship and knowing across communities can contribute to the theory and practice of information system design. Through the development of learning environments that span communities and support engagement of practice and appropriate modes of participation, organizations can create, focus, and transfer system knowledge through an apprenticeship model of learning. For information systems researchers, the model provides new opportunities for examining and evaluating requirements definition approaches. All of the issues raised in this article require further investigation. Many opportunities exist for future research to further understand, validate, and extend this work.

For practitioners, the model provides insights into the importance of context in understanding what is known in organizations and the importance
of practice in forming the way that we know it. For project managers and executives, creating an environment in which both designers and users are encouraged to be both masters and apprentices may well prove to be a critical factor for project success.

REFERENCES


Bryant, A. (2000). *It’s engineering Jim … but not as we know it*. Software engineering: Solution to the software crisis, or part of the problem? In *Proceedings of the 22nd International Conference on Software Engineering* (pp. 78 – 87), Limerick, Ireland.


internal technology transfer. *Academy of Management Journal, 36*(5).


*This work was previously published in the International Journal of Technology and Human Interaction, edited by B. C. Stahl, Volume 3, Issue 4, pp. 15-32, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 6.6
Grounding Business Interaction Models:
Socio–Instrumental Pragmatism
as a Theoretical Foundation

Göran Goldkuhl
Linköping University, Sweden, & Jönköping International Business School, Sweden

Mikael Lind
University College of Borås, Sweden, & Linköping University, Sweden

ABSTRACT

In the information systems field there exist several theories for guiding the evaluation and design of information systems. These theories need to be transparent and harmonious. In this chapter, business action theory (BAT) as a domain ontology for business interaction and business processes is clarified by elaborating on socio-instrumental pragmatism (SIP) as a base ontology. SIP is an eclectic theory synthesizing several pragmatic theories from reference disciplines outside the IS area. One purpose of SIP is to enable seamless theorizing in the IS area. In this chapter we put forward the foundations of BAT and SIP which are then followed by grounding BAT in SIP. This grounding means that there will be an ontological clarification of BAT by specifying the social action and interaction character of business interaction.

INTRODUCTION

Business action theory (BAT) is an ontology and a practical theory for business interaction and business processes. During the last 10 years we have been working actively to continually improve business action theory. This knowledge evolution process can be characterized as empirically-driven theory development. The goal has been to create an empirically, internally, and theoretically grounded theory for business interaction. Today, BAT has the epistemological status of a multi-grounded theory (Goldkuhl & Cronholm, 2003; Lind & Goldkuhl, 2006). In this chapter we will
further ground BAT as a business domain ontology in a higher level domain ontology (Guarino, 1998) for the distinction between different types of ontologies.

The basic characteristics of a practical theory have earlier been elaborated by Cronen (1995, 2001) and Craig and Tracy (1995). Cronen (1995, p. 231) describes practical theories in the following way:

_They are developed in order to make human life better. They provide ways of joining in social action so as to promote (a) socially useful description, explanation, critique, and change in situated human action; and (b) emergence of new abilities for all parties involved._

Practical theories should help us to see things, aspects, properties, and relations which otherwise would be missed (Cronen, 2001). The constituents of a practical theory have lately been elaborated by Goldkuhl (2006). Goldkuhl emphasizes conceptualizations, patterns, normative criteria, design principles, and models as (partially overlapping) such constituents of a practical theory. BAT is today regarded as a practical theory since all these constituents have been elaborated. For further elaboration, see Goldkuhl (1996), Lind and Goldkuhl (2003), and Goldkuhl and Lind (2004). The choice for us focusing on BAT is based on our good experiences in adopting BAT in practical situations and due to shortcomings in other theories for business interaction (Goldkuhl, 2006).

The notion of ontology is in this chapter was conceived as a particular system of categories accounting for a certain vision of the world (Guarino, 1998). Essential in the conception of ontology is conceptualization, which can be made on more or less formal foundations. According to Gruber (1993), an ontology is to be conceived as a specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects. Guarino (1998) claims that an ontology is a logical theory accounting for the intended meaning of a formal vocabulary, but not necessarily that the formal vocabulary is to be a part of a logical language, as, for example, it may be a protocol of communication between agents. Different conceptualizations of the world thus need to be included in the ontology. BAT as a practical theory is based on a pragmatic paradigm that sees scientific knowledge as means to improve human practices (Dewey, 1938). This also means a special interest in social actions constituting the world. See, for example, Goldkuhl (2005a), building on Mead (1938). The characteristics of a practical theory and requirements to put upon the domain-dependent ontology as BAT thus strongly overlap.

The starting point for the development of BAT was when Goldkuhl (1996), at the first language action perspective for communication modeling (LAP) conference, criticized the action workflow model of Medina-Mora, Winograd, Flores, and Flores (1992) for being asymmetric. See Goldkuhl (1996) for the complete criticism. An alternative to the action workflow model was presented which was the first version of the BAT model. This model emphasized business interaction as an exchange process with mutual commitments, fulfilments, and satisfaction. Both action workflow and BAT were founded in the language-action perspective (Winograd & Flores, 1986). This means that there were many theoretical affinities between the two models. There were, however, also substantial differences.

The BAT-model has since then been applied in many action research projects concerning co-design of business processes and IT (e.g., Axelsson, Goldkuhl, & Melin, 2000; Goldkuhl & Melin, 2001). Due to experiences from these applications of the model, it was continually redeveloped (Goldkuhl, 1998; Goldkuhl & Lind, 2004). Some essential characteristics of the BAT-model are that (see the section on BAT for a more thorough description) it emphasizes:
Grounding Business Interaction Models

- Business interaction between customer and supplier as two actor roles; supplier and customer
- A number of different exchanges, constituted by communicative and/or material acts organized as patterns of initiatives and responses
- The continual development of business relations and business capabilities
- The acknowledgement of other parties in the value transformation context
- The recurrence between several following business transactions
- The difference between interaction with the potential and the particular customer

The BAT-model originated within the language-action perspective with its basis in speech act theory (e.g., Searle, 1969; Habermas, 1984). However, from the start there was a criticism towards a narrow focus on only speech acts in business interaction (Goldkuhl, 1996). The fulfilments phase in the BAT model as a value exchange have been emphasized from its inception. There have been other kinds of deviation from a pure LAP orientation. The use of predefined communication patterns has been criticized by Goldkuhl (2003).

The BAT-model is built on several foundational concepts as roles, phases, and actions. One important attempt to clarify the BAT ontology was made in Lind and Goldkuhl (2003) where five conceptual layers were described. These layers were (from bottom to top): business act, action pair, exchange, business transaction, and transaction group. Through this work the need to ground this ontology not only in LAP was obvious. Other social action theories were used as basis.

Parallel to the development of BAT was the evolution of socio-instrumental pragmatism (SIP) as a foundational and progenitive theory for several practical theories in the IS area (Goldkuhl, 2002; Goldkuhl & Ågerfalk, 2002). By progenitive, we mean that SIP gives a theoretical grounding to other theories. SIP is in this chapter regarded as a domain ontology on a higher level than BAT; a base ontology since it describes general social domain concepts, which are independent of a particular problem or domain. For some ontology-researchers SIP would probably be regarded as a social domain ontology. Due to the need to distinguish between BAT as a business domain ontology (Guarino, 1998) from SIP as a social domain ontology, we characterize SIP as a base (domain) ontology. Goldkuhl and Röstlinger (2003) made an articulation of SIP as foundational basis for workpractice theory (Goldkuhl & Röstlinger, 2003), information systems actability theory (Goldkuhl & Ågerfalk, 2002), and sociopragmatic communication theory (Goldkuhl, 2005b) besides BAT. SIP is an eclectic theory synthesizing several pragmatic theories from reference disciplines outside the IS area (Goldkuhl, 2005a). One purpose of SIP is to enable seamless theorizing in the IS area (Goldkuhl, 2005).

Although important steps have been taking by Goldkuhl and Röstlinger (2003a) and Lind and Goldkuhl (2003) in grounding the BAT model in socio-instrumental pragmatism, much work remains. The purpose of this chapter is to clarify BAT as a business interaction ontology (business interaction as the domain) by elaborating on SIP as a base ontology for such clarification. Different concepts of business interaction will be grounded in foundational concepts of social interaction as they come through in socio-instrumental pragmatism. This will lead to three types of theoretical contributions (ontology clarifications):

- Clarification of business interaction concepts in the BAT-model
- Clarification of social interaction concepts within socio-instrumental pragmatism
- Conceptual grounding of the BAT-model in socio-instrumental pragmatism (conceptual relating)
After this introduction, a more thorough description of the BAT-model will follow. This description also includes, for the purpose of putting the BAT-model on the scene of contemporary concepts, relationships towards other similar concepts will be explicated. Following this description we will put forward the foundations of the socio-instrumental pragmatism as the base ontology used in this chapter. After these two foundational parts of this chapter, different business interaction phenomena according to BAT will be grounded in socio-instrumental pragmatism. The chapter will conclude with some highlights from this grounding.

**BUSINESS ACTION THEORY**

**Overview of the BAT Model**

The BAT-model is a phase-model describing a generic business interaction logic (Goldkuhl, 1998).

*Figure 1. Levels of business interaction (Goldkuhl & Lind, 2004)*
Grounding Business Interaction Models

BAT describes customer—supplier interaction. Business interaction should be understood as the performance of business acts (Lind & Goldkuhl, 2003). A vital aspect here is that a business act can be both a so-called speech act (a communicative act) and a material act. Thus business interaction is not restricted to communication but also comprises vital material acts as the exchange of products vs. money.

The BAT-model is a generic framework for business dyads. Being a generic framework means that it covers different types of business interaction as B2B (business-to-business) and B2C (business-to-consumer). It also covers different types of products, both goods and services and both standard and tailor-made products. The application scope of the model is to be used for evaluation, modeling, and design of business interaction. In the current version of BAT (Goldkuhl & Lind, 2004), two levels of business interaction are distinguished: the market level and the dyadic level (see Figure 1). On the market level, suppliers and customers search for knowledge and contacts concerning the correspondent party. The interaction on this level is, according to BAT, driven by a general business interest of both suppliers and customers.

When a contact is established between a particular supplier and a particular customer, the general business interest is turned into a particular business interest. The business interaction moves to the dyadic level where business transactions take place. According to the BAT-model a business transaction consists of four phases: (1) proposal phase, (2) commitments phase, (3) fulfilments, and (4) assessments phase (Figure 2). Each phase consists of exchanges of a particular character, for example, the first phase consists of exchanges of proposals.

The BAT-model is characterized as a comprehensive framework (Goldkuhl & Lind, 2004) that sees business action as a building block, emphasizes the exchange character of business interaction, adopts a symmetric view on business parties and their interaction, and acknowledges both communicative, material, and financial interaction.

Essential Concepts

The BAT-model is based upon the business action as the basic unit of analysis. Conditions for and results of business actions are action objects that could be of communicative and/or material character. The notion of business acts builds upon the notion of social action. Performing social actions—either communicative or material—introduces relations between two actors in roles (supplier and customer). The performance of a communicative action (like a business order) introduces clearly certain relations between sender and recipient (Habermas, 1984). Based upon the generic model of social action (see the next section) a business act is defined as “performance of a communicative and/or material act by someone aimed towards someone else” (Lind & Goldkuhl, 2003, p. 335). Business acts are often multifunctional. One example of multifunctionality is that the order both represents a request to the supplier to deliver something and a commitment of paying for the delivery corresponding to the order. Another example of multifunctionality is that a delivery of a product can both be a change of place of some material stuff and a fulfilment of a request and a promise.

The interactivity in the BAT-model is constituted by grouping two business acts into an action pair. The basis for grouping business acts into an action pair is that one business act functions as a trigger for another act, which will have the function of a response. Action pairs are patterns of triggers and responses. One example in business interaction is an order as the trigger and an order confirmation as the response of that trigger. By issuing an order the interventionist (i.e., the customer) expect the recipient (i.e., the supplier) to respond in a certain way: by confirming the order, by negotiating, or by turning the order away.
Action relations between the actors are created through such an interaction.

One or several action pairs, that is, patterns of action pairs, constitute exchanges between actors. An exchange means that one actor gives something in return for something given by another actor. Exchanges are, however, not only related to exchange of value (such as physical goods in return for money)—exchange is related to different kinds of communicative as well as material business acts. As indicated in the description of the BAT-model, exchanges are of different types (such as proposals, commitments, values, and assessments). Exchanges constitute the core of business interaction. Each type of exchange distinguishes a business phase.

Several exchanges together constitute a business transaction. There is a certain logic of interaction between the supplier and customer when doing business. This logic is the pattern covered by the business transaction (Axelsson et al., 2000; Goldkuhl, 1998). A business transaction has different states (in the social and material world) where the initial state is that the customer has a need and the supplier has a corresponding ability. The interaction between the two parties is determined by their dispositions, such as capabilities, business interests, and needs. By going through a number of phases consisting of exchanges the goal is to arrive at a state where both supplier and customer have satisfied (parts of) their needs. Patterns of exchanges, covering different phases, between the two roles continually evolve relations between the two business parties.

**BAT in Relation to Other Business Interaction Frameworks**

Within the language/action (L/A) community there is an interest for business interaction. The strength of the language/action perspective is that it is based on the idea that communication is not just transfer of information. When you communicate you also act (Searle, 1969). Actions are performed, including building commitments and agreements between business parties. Agreements are to be regarded as the backbone of L/A approaches. Both agreements on what to do and agreements on performed actions are accentuated. Such emphasis on agreements causes a division of the communication process into three or four phases.

Action workflow (AW) (Medina-Mora et al., 1992) and DEMO (Dietz, 1999) are two frameworks founded in the language/action tradition. Although valuable features exist in these frameworks, certain limitations also exist. The interactive character is not fully developed. See Goldkuhl and Lind (2004) for further explanation of this criticism. As a reaction towards these deficiencies, the BAT-model was introduced (Goldkuhl, 1996, 1998; Goldkuhl & Lind, 2004; Lind & Goldkuhl, 2003). The BAT-model has been applied in different studies (e.g., Axelsson et al., 2000; Goldkuhl & Melin, 2001; Lind, 2002).

As mentioned in the introduction, there is a need for comprehensive frameworks covering business interaction. The BAT-model is claimed to be one. Another promising attempt to create such comprehensive framework has been made by Weigand and van den Heuvel (1998) in which metapatterns for electronic commerce is proposed. This framework is built on the idea to integrate different L/A-oriented approaches, such as DEMO, action workflow, and BAT for business modeling.

Originally outside the language/action tradition, Schmid and Lindemann (1998) have presented a reference model for electronic markets that in later works (Lechner & Schmid, 2000) has been expanded to a more general framework—a media reference model (MRM). A comparison of BAT and MRM has been made in Petersson and Lind (2005).
SOCIO-INSTRUMENTAL PRAGMATISM

Social Action

Socio-instrumental pragmatism (SIP) is a foundational theory (base ontology) specifically aimed for studies concerning information systems (e.g., Goldkuhl, 2002, 2005a; Goldkuhl & Röstlinger, 2003). It is an eclectic synthesis based on several action-theoretic frameworks form reference disciplines, like, for example, philosophy, linguistics, sociology, psychology, and organization theory. It is based on foundational pragmatic insights leading to action as a core concept. Herbert Blumer, one of the founders of symbolic interactionism, claims that

...the essence of society lies in an ongoing process of action—not in a posited structure of relations. Without action, any structure of relations between people is meaningless. To be understood, a society must be seen and grasped in terms of the action that comprises it. (Blumer, 1969, p. 71)

The social world is created and recreated through human actions. This means that most actions are of social character. The great sociologist Max Weber has made a classical definition of social action: “That action will be called social which in its meaning as intended by the actor or actors, takes account of the behaviour of others and is thereby oriented in its course” (Weber, 1978, p. 4). Our interpretation of this definition is that a social action (performed by an actor) has social grounds (“takes account of the behaviour of others”) and social purposes (“thereby oriented in its course”).

From this follows the basic idea in SIP that (most) actions are directed towards other humans. There are addressees of most actions. When we, as human actors, create or change some material object, there may be addressees for this action object. When we say something, there are definitely addressees for these communicative actions. In SIP, there is a basic model of social action (Figure 3). This model consists of two actors. One actor is conducting an intervening action (a communicative action or a material action) directed towards

Figure 3. Socio-instrumental action: A basic model (Goldkuhl, 2005a)
the addressee. The addressee performs a receiving action, that is, the receipt of a material object or the interpretation of a message. The intervening actor is the focused actor in the model. This actor has social grounds for the action. The actor pre-assesses external and internal grounds in a deliberative phase before the intervening action. After intervention, the actor postassesesses the result and the effects. This builds on a continuity model of actions with a division into three stages: (1) pre-assessment, (2) intervention, (3) post-assessment (Goldkuhl, 2004). Originally, this builds on Mead’s (1938) four stage model of human action: impulse, perception, manipulation, and consummation. In SIP these stages have been renamed and the first two stages have been integrated into one.

In SIP, the action is seen as knowledgeable. The actor uses her knowledge in order to make a difference in the external world (Dewey, 1931). The actor is seen as purposeful with abilities to reflect and deliberate before intervention (Dewey, 1931; von Wright, 1971). The SIP model of human action also builds on an important difference between result and effect (von Wright, 1971). The result is what the actor produces, that is, the direct result of the action. The result is what lies in the range and control of the actor. Effects are what arise as consequences of the performed action.

Social action takes place in a temporal and spatial context. We can distinguish between an intratemporality and an intertemporality of an action. With intratemporality we delimit to the actor (an intra-actor characterization). This means the phasing of pre-assessment, intervention and postassessment described above. The intertemporality emphasises the actor interacting with other actors: In this case we follow the division of (1) social grounds for action, (2) the (interventive) action, and (3) the social effects of action.

The SIP model also acknowledges that human actions often are mediated through the use of some instrument (e.g., Vygotsky, 1962). Instruments are mediational means for human action with enabling and constraining functions for the actor. Many actions can be seen as transformation of objects. A base object is transformed into a result object (Goldkuhl & Röstlinger, 2003).

Social Interaction and Relations

As every model, Figure 3 is a simplification. It describes a one-way action from the focused actor to the addressee. There is no explicit interaction depicted in this model. Within the social grounds, there may be an implicit interactional component. There may be, for example, a request of some action from the addressee actor in the figure. This basic social action model is therefore complemented with a social interaction model (Figure 4). In this model the interactive nature of social life is more explicitly depicted. The interaction model describes two related actions. One action is an initiative for the subsequent action, which is seen as a response to the first. The response action can then function as an initiative for subsequent responsive actions. The concepts of initiative and response is fetched from the concept adjacency pair in conversation analysis (Sacks, 1992) and later refined in dialogue theory (e.g., Linell, 1998). Goldkuhl (2005a, 2005b) describes the integration and use of these concepts in socio-instrumental pragmatism.

Another important aspect of social interaction is also depicted in Figure 4. Social actions imply relational changes between actors. When one actor says something or does something else directed to an addressee, certain social relations between the actors are established. For example, when an actor is posing a question, expectations are arisen that an answer or some other adequate response will follow. SIP builds on communication theories where this kind of relational changes are emphasized; speech act theory/communicative action theory (Habermas, 1984; Searle, 1969) and conversation analysis (Sacks, 1992). Some scholars (e.g., Linell, 1998) conceive these theories as contradictory. In SIP, these theories are conceived as complemen-
tary. Goldkuhl (2003, 2005b) has described how aspects from speech act theory and conversation analysis are integrated into one congruent communication model founded in SIP.

What is a social relation? What does it mean? Where do social relations exist? In order to discuss this matter, we need to use the concepts of a focused actor (“ego”) and another actor (“alter”). Relations exist as ego’s apprehension of the alter and vice versa. In order to succeed in social interaction, parts of these relational apprehensions need to be intersubjective and not only intrasubjective (these concepts will be further discussed). Different emotions concerning each other may be personal and thus intrasubjective. However, in order to conduct a business conversation in a mutually understandable and coordinated way, different relations between ego and alter need to be intersubjective to a sufficient degree. If ego asks for a product price, then expectations are raised for some answer from alter. After this question has been posed, both parties are usually aware of the expectations for a proper response (Heritage, 1984). Otherwise there will be a breakdown in the communication. Every “move” in the social interplay of a conversation introduces new relations and changes old ones. A communicative action is usually both a response to previous actions and an initiative to subsequent actions. This means that a certain action may change earlier relations and introduce new ones. This is described in Table 1 and Table 2.

Table 1 describes the different types of relational development that occur from ego action. The table both pinpoints the ego action in relation to alter action as well as ego actions as response (to earlier actions) and projection (for future actions). As stated, an ego action has, as an initiative, connections to future actions. In a communicative act, there are often projections of possible future actions (Goldkuhl, 2005b). An ego action gives expectations for future alter actions. A request is a typical example; the request gives expectations for some future alter actions. An ego action may not only have connections to alter future actions, but also to ego coming actions. A promise is a typical example of introducing commitments for future actions (Searle, 1969). An ego action has, as a response, connections to earlier actions. This relational aspect of an ego action means that it is a reply to alter’s earlier action and the kind of expectations which are raised in this alter ac-
We call this relational aspect replication. One example of this is ego’s answer to an earlier posed question by alter. An ego action may not only have connections to alter previous actions, but also to ego own actions. Ego may have done some commitment in a previous action (e.g., a promise), and the present ego action may be in accordance or not to this commitment. Being in accordance, the ego action is a fulfilment and if not, it will be a deviation.

Table 1 described the relational meanings of an ego action. Table 2 describes the relational meanings of an alter action. Here we use the postassessment perspective explicitly. The situation described is when ego makes a (post-)assessment of alter’s conducted action. What relational changes may occur after alter’s action? If there are any action projections made in alter action, relational changes of commitment or expectations may occur. If alter requests something to be done by ego, then an expectation (towards ego) is introduced. If alter promises to do something, commitments for future actions are introduced.

What alter does can also be assessed in relation to ego previous actions and alter previous actions. If alter replies to a question earlier posed by ego, then different kinds of experiences may arise. If alter gives a proper answer then ego may conclude that alter seems to be knowledgeable person. If the answer does not seem to be adequate or correct, then ego’s experiences of alter as knowledgeable may diminish. Or ego may reflect upon ego’s question (earlier action), that it was not so properly formulated. This follows the principle of action reflexivity described by, for example, Giddens (1984). All actions have repercussions back on the actor himself. Table 1 and Table 2 have been introduced here as an important refinement of SIP and will be used when performing a SIP analysis of business actions.

### Realms of the World

With socio-instrumental pragmatism comes also a division into different realms of the world. Goldkuhl (2002) has made a principal division into
different ontological categories (realms). These are depicted in Figure 5. The different ontological categories are the humans, their inner worlds, their actions, and the external world consisting of signs, material artefacts, and nature. Actions are subclassed into three types: (1) Intervening actions, that is, making changes in the external world, (2) interpreting actions, that is, making sense of the external world, and (3) reflective action, that is, making internal changes through thought processes. This follows the division made by Schutz (1970) in overt actions (intervention) and covert actions (interpretation and reflection). Signs and material artefacts are created or changed through intervening actions.

Human inner world is divided into an intra-subjective part and an intersubjective part. The intersubjective part is knowledge that is shared with fellow-actors, for example norms, rules, linguistic competence and methods. This intersubjective part is known under different labels as, for example, social institutions (Berger & Luckmann, 1966; Scott, 1995), structure (Giddens, 1984), and culture (Duranti, 1997).

The SIP consists of some basic ontological building blocks as described in Figure 5. More complex phenomena are built from these elements. One very important phenomenon of complex nature in business interaction is organizations. An organization (a company), acting as a supplier or as a customer is of prime interest for business interaction models. What is an organization in a SIP view? Goldkuhl and Röstlinger (2003) has presented such an ontological clarification based on socio-instrumental pragmatism. “An organization exists as an agreement (a communicative

Figure 5. Different realms of the world (Goldkuhl, 2002)
fact) between the principals and other parts of the society. Through such constitutive actions, an organization is given a formal authority to act” (Goldkuhl & Röstlinger, 2003, p. 157). An organization is the result of constituting communicative acts between its principals and legal authorities. As such it is an intersubjective phenomenon depending on formal communicative acts. In a continual existence an organization must comprise different resources; financial and material resources and employed humans and their knowledge. In SIP, organizations are seen as actors (Goldkuhl and Röstlinger, 2003; Taylor & Van Every, 2000). They can, however, not act by themselves. Organizations act through their human agents and sometimes also through artificial agents, like IT systems. Organizational agents (like employees) act as representatives of the organization.

GROUNDING BUSINESS INTERACTION ACCORDING TO BAT IN SOCIO-INSTRUMENTAL PRAGMATISM

Socio-instrumental pragmatism has contributed to sharpen BAT in several respects. Lind and Goldkuhl (2003) and Goldkuhl and Röstlinger (2003) are two important earlier contributions. Further ontological clarifications are needed in order to specify the social action and interaction character of business interaction. This will be shown below.

BAT is a theory and model for business interaction. It is built from some basic concepts as

- Actors in roles (supplier and customer) can be a human actor or an organization
- Business actions (e.g., offering, ordering, delivering, and complaining) with corresponding action objects of communicative (e.g., offer, order, complaint) or material (e.g., goods) character
- Business phases consisting of actions concerning similar type of exchange (i.e., meaningful action aggregates)
- Business transactions consisting of actions within business phases from proposal to assessment (i.e., meaningful action aggregates)
- Relations between business parties which continually change through the business transaction (i.e., changes in the intersubjective world); a certain role plays the business agreement established in phase 2 and resolved in later phases
- Actor dispositions, which are preconditions for interactive business actions; for example, capabilities (of human and artefact character), business interests, and needs (human inner world and often expressed in signs)

Every generic business action described in the BAT-model is a social action and can be described in accordance with the generic model of social action (Figure 3). One actor (business party) conducts an action (e.g., offering) directed towards the other actor. An action object (e.g., an offer) is created/presented to the other actor. This other actor receives the action object and interprets it. Through this interpretation subsequent actions can be performed (e.g., ordering). This means that business actions are performed as a social interplay. One action functions as an initiative to another action, which is a response to the first action.

We will below perform a social action analysis of some generic business acts from the four business phases. In this analysis, we use concepts and perspectives from SIP. This will lead to contextual definitions of business actions emphasizing their social interaction character. When clarifying the business actions, the performing actor and the addressee must, of course, be described. This is done together with stating the action direction (from intervening actor to addressee). Social grounds for the action are also important for describing the
action. The actor should not be seen as a social marionette just functioning in a stimulus-response fashion. The actor has capabilities for deliberation before action. SIP is based on speech act theory (e.g., Searle, 1969) with the important differentiation into illocutionary force (here called action modus) and propositional contents. The focused action’s relations to future actions are important. They are, however, not described in this type of table. This aspect is instead described in special relational tables below, based on the principal Table 1 above.

The social action analysis of some of the generic business actions are thus a step towards defining key concepts and intentionally relating them to each other within the domain ontology (BAT). Relations between different concepts are to some extent elaborated on by using tables. An

### Table 3. Characterization of actions in proposal phase

<table>
<thead>
<tr>
<th>Action</th>
<th>Action direction</th>
<th>Social grounds</th>
<th>Deliberation</th>
<th>Action modus</th>
<th>Propositional contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inquiry</td>
<td>Customer → Supplier</td>
<td>Knowledge about possible suppliers</td>
<td>Needs for product; exploration desire</td>
<td>Product question (information seeking)</td>
<td>Needs → products? Product → product details, delivery terms?</td>
</tr>
<tr>
<td>Invitation to tender</td>
<td>Customer → Supplier</td>
<td>Limited knowledge about suppliers</td>
<td>Needs for product; desire for (alternative) tenders</td>
<td>Purchase interest</td>
<td>Type of product, desired commercial terms</td>
</tr>
<tr>
<td>Offer</td>
<td>Supplier → Customer</td>
<td>Knowledge about possible customers and their needs</td>
<td>Desire to make a competitive offer</td>
<td>Sales offer (proposal)</td>
<td>Product, product characteristics</td>
</tr>
<tr>
<td>(Counter) Bid</td>
<td>Customer → Supplier</td>
<td>Knowledge about suppliers and their products and offers</td>
<td>Precise product needs, business consideration</td>
<td>Purchase proposal</td>
<td>Product, product characteristics, commercial terms</td>
</tr>
</tbody>
</table>

### Table 4. Relational characterization of offer

<table>
<thead>
<tr>
<th>Supplier actions (focused actor)</th>
<th>Responded (earlier actions)</th>
<th>Projected (future actions)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>Commitment to sell in accordance with given offer</td>
</tr>
<tr>
<td>Customer actions (other actor)</td>
<td>Replication to inquiries, tender, invitations, counterbids</td>
<td>Expectation to order product</td>
</tr>
</tbody>
</table>
important task in ontology declaration is to make intentional relations (cf. Guarino, 1998) explicit. This task has been initiated in our former work (e.g., Lind & Goldkuhl, 2003). The more explicit definitions put forward in this chapter should be seen as a further extension of this work.

In Table 3, we have described four types of actions from the proposal phase. Offer is the typical action of a supplier in this phase. We have described three types of customer actions: inquiry, invitation to tender, and counter bid. Table 3 is followed by Table 4, in which we have made a relational characterization of the offer action based on the relational development through the supplier’s actions.

**Contractual Phase**

The second phase in the BAT-model is the commitments phase. In this phase the business parties arrive at a business contract. The two main actions in this phase are the customer ordering and the supplier confirming. These two actions are characterized in Tables 5-7. We have also described the supplier’s invoicing as an action in this phase (Table 5). It might be seen as disputable. In the first BAT-model (Goldkuhl, 1996), it was described as part of the fulfilment phase. We have classified the invoicing to be part of the contract phase since these actions are concerned with directives for the exchange of value in the fulfilment.

<table>
<thead>
<tr>
<th>Action</th>
<th>Action direction</th>
<th>Social grounds</th>
<th>Deliberation</th>
<th>Action modus</th>
<th>Propositional contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order</td>
<td>Customer → Supplier</td>
<td>Given offers</td>
<td>Evaluation of offers vs. needs</td>
<td>Purchase order</td>
<td>Product, no of items, commercial terms</td>
</tr>
<tr>
<td>Delivery confirmation</td>
<td>Supplier → Customer</td>
<td>Order</td>
<td>Checking delivery possibilities</td>
<td>Delivery promise</td>
<td>Product, no of items, delivery terms</td>
</tr>
<tr>
<td>Invoice</td>
<td>Supplier → Customer</td>
<td>Delivery agreement, planned or conducted delivery</td>
<td>Appropriate time for payment</td>
<td>Payment directive</td>
<td>Delivery, payment amount, payment details</td>
</tr>
</tbody>
</table>

**Table 6. Relational characterization of order**

<table>
<thead>
<tr>
<th>Customer actions (focused actor)</th>
<th>Responded (earlier actions)</th>
<th>Projected (future actions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication to offer</td>
<td>Commitment to pay for future delivery</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Supplier actions (other actor)</th>
<th>Responded (earlier actions)</th>
<th>Projected (future actions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expectation to confirm and deliver products in accordance with order</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
phase. The order is a directive for delivery. The invoice is a corresponding directive for payment. Table 6 covers the relational characterization of the order action performed by the customer (as the ego). In Table 7, the relational characterization of the delivery confirmation action.

**Fulfilment Phase**

The main actions in the fulfilment phase are the delivery and the payment. These actions of exchange of value are mainly material and not communicative. Therefore we have left out the column “propositional contents” in Table 8.

**Assessment Phase**

The assessment phase can consist of positive and negative assessments. The two negative assessments, complaint and reminder to pay, are described in Table 9. A relational characterization of the complaint action is made in Table 10.

**CONCLUSION**

The purpose of this chapter has been to clarify BAT as business interaction ontology (a domain ontology). For this purpose socio-instrumental pragmatism (SIP) has been used as a base ontology. SIP uses social action as the core concept. A social action, which can be communicative and/or material, means that the action is conducted and directed towards another actor. Social actions have social grounds and social purposes. To act towards another actor in patterns of interaction implies relational changes between the actors. Every move in a social interplay introduces new

---

**Table 7. Relational characterization of delivery confirmation**

<table>
<thead>
<tr>
<th>Supplier actions (focused actor)</th>
<th>Responded (earlier actions)</th>
<th>Projected (future actions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fulfilment of given offer</td>
<td>Commitment to deliver products</td>
<td></td>
</tr>
<tr>
<td>Replication to order; acknowledging the order leading to an establishment of a contract</td>
<td>Expectation to pay for delivered products</td>
<td></td>
</tr>
</tbody>
</table>

**Table 8. Characterization of actions in fulfilment phase**

<table>
<thead>
<tr>
<th>Action</th>
<th>Action direction</th>
<th>Social grounds</th>
<th>Deliberation</th>
<th>Action modus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delivery</td>
<td>Supplier → Customer</td>
<td>Business contract (order + confirmation)</td>
<td>Time for delivery</td>
<td>Product provision</td>
</tr>
<tr>
<td>Payment</td>
<td>Customer → Supplier</td>
<td>Business contract (order + confirmation), invoice, possibly conducted delivery</td>
<td>Time for payment</td>
<td>Money transfer</td>
</tr>
</tbody>
</table>
Grounding Business Interaction Models

Table 9. Characterization of actions in assessment phase

<table>
<thead>
<tr>
<th>Action</th>
<th>Action direction</th>
<th>Social grounds</th>
<th>Deliberation</th>
<th>Action modus</th>
<th>Propositional contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complaint</td>
<td>Customer → Supplier</td>
<td>Delivered products</td>
<td>Discontent with product use</td>
<td>Express of discontent, request for improved product</td>
<td>Product, product deficiencies</td>
</tr>
<tr>
<td>Reminder to pay</td>
<td>Supplier → Customer</td>
<td>Delivered products, absent/ unsatisfying payment</td>
<td>Discontent with payment</td>
<td>Claim for payment</td>
<td>Delivery, absent payment</td>
</tr>
</tbody>
</table>

Table 10. Relational characterization of delivery of deficient products (grounds for complaint) from customer viewpoint

<table>
<thead>
<tr>
<th></th>
<th>Responded (earlier actions)</th>
<th>Projected (future actions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer (ego actor)</td>
<td>Experiences (ego actions): Could have chosen a better supplier</td>
<td>Expectation to customer: To give feedback concerning product use</td>
</tr>
<tr>
<td>Supplier (alter actor)</td>
<td>Experiences (alter actions): Unreliable supplier</td>
<td>Commitment of supplier: Products should be useful</td>
</tr>
</tbody>
</table>

relations and changes old ones. It is thus important to consider several adjacent (social) actions, communicative as well as material, in patterns of interactive acts between the two (or several) parties. The BAT-model emphasizes phases and interaction patterns for establishing expectations, fulfilment of, as well as evaluation of fulfilled expectations.

In order to clarify BAT as business interaction ontology we have in this chapter made SIP-based social action analyses of the four business phases constituting the BAT-model. This has been done by stating action direction, social grounds, deliberation, action modus, and propositional content for essential business acts within each of the four business phases. For each of the main actions in each business phase, relational characterizations has also been made. In order to make such characterization, we have, in this chapter, introduced the notion of ego as the focused actor and alter as the other actor. In the social interplay the two actor roles, supplier and customer, alternate between being an ego or an alter. In each phase of the business interaction sequence possible business actions are determined ego’s response to earlier actions and projected actions towards the alter.

This analysis has led to a more explicit grounding of the BAT-model in SIP as an ontological
foundation. The BAT-model and SIP co-exist in the same sociopragmatic “family” of theories with SIP on a more foundational level. This pursued analysis has also implied a further sharpening of some social interaction concepts within socio-instrumental pragmatism. The typologies of different kinds of relations in social interaction are such a theoretical result.

The work reported in this chapter is to seen as a further step towards a practical theory. SIP is a foundational ontology building on social action. It can thus now be claimed that BAT has been further grounded in general social action concepts. This work also means that BAT, as a practical theory, now should be seen as a more congruent domain ontology—at least within the four phases of the BAT-model focused in this chapter.

BAT is intended to be used as a pragmatic instrument for understanding business interaction. It is a comprehensive framework to be used for several purposes. It can be used as a conceptual instrument when evaluating existing business interaction. The framework can also be used for modeling and designing business interactions. It reminds the designers about different dynamic features of business interaction as exchanges, recurrence, evolvement of business relations, and capabilities. It can give structure in the design process and be a basis for the important design of business actions and the allocation of actions to different agents (human agents and IT artefacts). Development of IT-support for business interaction is to be seen as a basic development of business capability.

The work reported in this chapter is to be seen as a rather informal metamodel (based on text and tables) pinpointing some key concepts in the BAT-model. An important step in the future is to make a more formalized metamodel based on the insights derived from grounding BAT in SIP, covering all key concepts in BAT.

Another step in future ontological clarification of the BAT-model is to investigate the relations between different business phases (initiatives and responses as the glue between different phases) as well as looking into different types of dyadic interactions. This is an important task since the logic of business interaction is conceived as establishing expectations (which is performed in two of the phases), fulfilling these established expectations and finally assessing the fulfilment. It is, therefore, important to highlight these relations since the actions in the latter phases are responses to earlier performed business actions. We also acknowledge a need to explicitly ground the differentiation between market-based interaction and dyadic interaction in the SIP ontology.

REFERENCES


*This work was previously published in the Handbook of Ontologies for Business Interaction, edited by P. Rittgen, pp. 69-86, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*

2089
Chapter 6.7
Managing Socio–Technical Integration in Iterative Information System Development Projects

Bendik Bygstad
Norwegian School of Information Technology, Norway

ABSTRACT

The challenges of information systems (IS) development have changed significantly in the past 15 years. From a situation where the main problem was to build stable systems based on a requirements specification, the IS project manager now faces challenges of integration; for example: how to integrate a new information system into large existing information infrastructures and unstable business processes? Building on a socio-technical perspective, four types of integration were identified and analyzed: external and internal stakeholder integration, and internal and external technical integration. A longitudinal case study of an e-business development project in the airline industry identified and highlighted some managerial challenges of integration. Findings include: Internal technical and (to some extent) stakeholder integration is well supported by traditional project management techniques and software engineering frameworks, such as the Rational Unified Process (RUP). However, the challenges of external stakeholder and technical integration is underrated in IS development research, and not well supported in RUP.

INTRODUCTION

Fifteen years ago, most IS projects were greenfield projects; on the basis of a requirements specification, a brand-new system was designed and programmed. When (eventually) finished, it was installed and run, and the users were trained in new interfaces and routines. The main challenges for the project manager were to get the requirements specification right and complete, and to develop the system to be technically stable.
Managing Socio-Technical Integration in Iterative Information System Development Projects

Of course, this is still a demanding craft, requiring specialized skills, sound methodologies and competent project management. However, the project manager today faces two more challenges:

• Most information systems are developed to support a business process, whether the scope is a dramatic process transformation or process improvement (Davenport, 1993; Laudon & Laudon, 2004). It is essential that the new system integrates with the business process, because a project that fails to do so will be a failure, even if the software product is well designed and programmed.

• A new system must also relate to an existing information infrastructure. The information infrastructure is seen as a heterogeneous network, comprising an installed base of technology, organization, culture and work practices (Hanseth & Lytinen, 2004; Hanseth & Monteiro, 1996). Both opportunities and constraints are heavily influenced by the attributes of this base. In a successful organization, this heterogeneous network is an immensely valuable resource; it constitutes the backbone of the organization. But in a world of change, it also may be a barrier to business adaptation or innovation, because the information infrastructure is hard and expensive to change.

The successful IS project manager must address both these challenges. In a way, this redefines the role of the project manager, making him or her an integrator of both social and technical forces. The question is: How is the manager going make people and technology work together in a complex and unstable setting?

Ideally, the process models of Business Process Reengineering (BPR) and IS development should answer this question. Unfortunately, the BPR community never really developed a full methodology for this integration (Giaglis, 1999).

During the late 1990s, the software engineering community embraced iterative and incremental process frameworks (Larman & Basili, 2003), such as RUP (Jacobson, Booch & Rumbaugh, 1999), Object-oriented Process, Environment and Notation (OPEN) (Henderson-Sellers & Unhelkar, 2000), Extreme Programming (Beck, 2000) and Dynamic Systems Development Method (DSDM) (Stapleton, 2003). The growing practice communities contend that these frameworks have the potential to integrate the business process and the new information system through step-wise stakeholder and technical integration (Blomberg, 2001; Kruchten, 2000; Stapleton, 2003). The basic mechanism is the short iteration that produces a small release that can be tested, integrated into the information infrastructure and assessed by the business organization.

This deserves attention not only in the realm of software engineering, but also in the broader context of information technology (IT)-based organizational change. While skeptics have pointed to lack of practical integration support for e-business architectures (Smolander, 2003), and poor project management support (Henderson-Sellers & Unhelkar, 2000), the large RUP community points to a number of success projects (Rational, 2006b). This paper explores, analytically and empirically, two important questions in this context:

• How can the IS project manager integrate the new information system with the business process and the information infrastructure?
• What integration support is there in software engineering frameworks, like RUP?

The rest of this paper is structured as follows. In the next section, the concept of integration is defined and discussed. Then, the integration support in RUP is briefly assessed. The research approach, longitudinal process research (LPR), is presented. Next, the integration challenge is
investigated empirically through a longitudinal case study of an IS project in the airline industry. The findings in the case study are discussed, and implications for practice and research are assessed. The last section concludes and points to further research.

DEFINING INTEGRATION

According to Webster, integration is “a combination and coordination of separate and diverse elements into a more complete or harmonious whole” (Webster’s, 1986, p. 1174).

Building on the socio-technical IS development tradition (Avison & Fitzgerald, 2003; Coakes & Elliman, 1999; Kling & Scacchi, 1982; Mumford, 1985) and actor-network theory-inspired IS research (Ciborra, 2000; Garrety, Robertson & Badham, 2004; Hanseth, 2002; Hanseth & Monteiro, 1996; Walsham, 1997), the two key element types to integrate are seen to be stakeholders and technology. Stakeholders are anyone (both humans and organizations) affected by the process, while technology may be networks, systems and components.

Stakeholder integration is defined here as a management activity that aims at facilitating cooperation between people and organizations, and helping them work together to create value. The mechanisms for achieving this may be structural (assigning formal roles and responsibilities) or process oriented, focusing on informal interaction (Heugens, van den Bosch & van Riel, 2002). In both cases, trust is important to keep the relationships alive.

Technical integration is, correspondingly, defined as a management activity to assemble components and systems in a way that make the resulting system work as intended. These components are often a mix of internally produced software, existing internal legacy systems, commercial software (COTS) and large networks, like Virtual Private Networks (VPNs) or the Internet. To make them work together is usually dependent on a combination of different actions: Designing correct interfaces, establishing stable operations and giving sufficient support from technical personnel. In the end, however, the most important element is probably correct use, thus connecting technical integration closely to stakeholder integration.

A basic idea in the socio-technical tradition is that the object of interest is not an artifact, a piece of technology, but rather the structuring and reproduction of a socio-technical network, consisting of both technical and social elements (Kling & Scacchi, 1982). Thus, the goal of an IS development project is not primarily seen as the construction of a software system, but to establish a working socio-technical network.

Both stakeholder and technical integration may be seen as management activities aimed at establishing linkages to different elements that have to work together. In large projects, the number of linkages will grow into complex networks, and this complexity leads to increased risk (Hanseth, 2002). The risk is enlarged by the fact that both the business process and the information infrastructure are unstable; they may change during the project time, due to external pressures or internal shifts of power.

Internal and external integration. Many integration activities are internal to the project, such as managing project teams, programming and testing components. But many integration activities are external to the project; aligning the project with business needs, modeling with future users, using external components and other systems. Thus, the integration process is multidimensional. It includes both stakeholders and technology, and the activities are both internal and external to the project. Thus, four types of integration are suggested in Table 1.

It is documented that integration is usually easier if done step-wise, because it reduces the technical risk and allows both the development project and the business process people to learn
Managing Socio-Technical Integration in Iterative Information System Development Projects

Early user participation has long been established as an important success factor (Garrity & Sanders, 1998). On a more political level, it also enables enrollment of important allies into the project (Hanseth & Monteiro, 1996). Early technical integration secures the integrity of the system; testing is performed in each iteration, and performance issues may also be addressed early (Jacobson et al., 1999). This is recognized in the iterative software engineering frameworks: For each iteration, new stakeholders are drawn into the project. And for each iteration, a small piece of solution is implemented, either in a test or pilot environment or in real production.

It may sound easy, but it is not. As Giaglis (1999) noted, the problem is not getting acceptance for the need of an iterative and integrative approach, but to actually manage it. The RUP framework was selected for this study as an example of a well-known framework. Therefore, the first step in this research was to assess to which degree RUP supports the IS project manager in his or her integration challenge.

ASSESSING INTEGRATION SUPPORT IN RUP

RUP is a large framework for software engineering developed by Rational Corp. in the late 1990s, building on Jacobson’s work at Ericsson (Jacobson et al., 1999). As illustrated in Figure 1, RUP is structured in four phases: inception, elaboration, construction and transition. Within each phase are one or several iterations consisting of disciplines, starting with business modeling and ending with the physical deployment of software components. RUP is both a process framework and a commercial Web-based product. The Web product is RUP online (Rational, 2006a), with 3,700 Web pages describing the principles of RUP and providing a rich toolbox of guidelines, risk lists and document templates.

To assess the degree of integration support in RUP, a rather simple but systematic analysis was

<table>
<thead>
<tr>
<th>Context</th>
<th>Stakeholder perspective</th>
<th>Technical perspective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integration within development process</td>
<td>(1) Internal stakeholder integration</td>
<td>(2) Internal technical integration</td>
</tr>
<tr>
<td>Integration with business process and infrastructure</td>
<td>(3) External stakeholder integration</td>
<td>(4) External technical integration</td>
</tr>
</tbody>
</table>

Table 1. Integration types

Figure 1. Structure of RUP (Rational, 2006a)
performed. The basic material was the templates of detailed activity lists provided in RUP online to help the IS project manager to plan the phases of the project. These lists are certainly not sufficient to plan a RUP project, but they give a reasonable picture of the profile of activities in each phase.

The assessment was done as follows: Each activity was classified within the four integration categories presented in Table 1. The resulting distribution of activities is shown in Table 2.

As the table shows, integration support is unevenly distributed. It is quite strong regarding internal technical integration throughout the phases, while it gives surprisingly little external technical integration support. Internal stakeholder integration is weak thoughout the project, confirming, perhaps, the critique against RUP for poor project management support (Henderson-Sellers & Unhelkar, 2000).

Support for external stakeholder integration is strong in the inception phase, but gets weaker in the later phases. A more detailed analysis of these activities shows that RUP gives strong high-level support to external stakeholder integration, but little practical support at the operational level. The same pattern is found regarding external technical integration. This may indicate another limitation in RUP: It endorses step-wise external integration, but little practical support.

### THE CASE STUDY: DEVELOPING AN E-BUSINESS SYSTEM FOR AN INTERNATIONAL AIRLINE

To investigate the challenges in more depth, a longitudinal case study was designed, studying an IS development project in the airline carrier industry. The research approach was LPR, which aims to study organizational change over time through intensive research in the actual context (Ngwenyama, 1998; Pettigrew, 1985, 1990). LPR focuses on building theories strongly embedded in the context of study. Reflecting Giddens’ structuration theory (Giddens, 1987), context is seen by Pettigrew not only as an environment, but as “a nested arrangement of structures and processes where the subjective interpretations of actors perceiving, comprehending, learning and remembering help shape the process. The processes are both constrained by context and shape contexts” (Pettigrew, 1990, p. 270).

Important criteria for data collection are, according to Ngwenyama (1998):

---

**Table 2. Distribution of RUP activities supporting the integration types through RUP phases**

<table>
<thead>
<tr>
<th>Phase type Integration</th>
<th>Inception</th>
<th>Elaboration</th>
<th>Construction</th>
<th>Transition</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Internal stakeholder integration</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>(2) Internal technical integration</td>
<td>11</td>
<td>23</td>
<td>23</td>
<td>14</td>
<td>67</td>
</tr>
<tr>
<td>(3) External stakeholder integration</td>
<td>14</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>(4) External technical integration</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>11</td>
</tr>
</tbody>
</table>
• Ongoing engagement with the research site, to observe changes over time
• Participant observation, to contextualize and make sense of observations
• Multiple data sources, to record different interpretations of events and ensure validity of findings

The case, a RUP development project at Scandinavian Airline System (SAS), was researched for 16 months, using several techniques for data collection: Project managers, developers and users were interviewed at four intervals. Project meetings were observed, and a vast amount of project documentation was analyzed.

Data collection was done in four phases, as summarized in Table 3.

Ngwenyama (1998) suggests three modes for data analysis: (1) Comprehensive analysis, which helps to identify underlying structures and patterns of the organizational process; (2) temporal analysis, which helps contextualize findings by placing events and situations in a narrative structure; and (3) member verification, which ensures that the case description and interpretation researcher are considered correct and meaningful to the organizational actors.

The SAS case was analyzed in the following steps: Data was coded following the guidelines of Miles and Huberman (1994). After the videotaped interviews were summarized and registered into an Atlas database, texts were coded with in-vivo codes, using only domain (project) terms. The large volume of project documentation was coded the same way.

First, a timeline with significant events was produced. Then, each iteration of the project was analyzed in detail, while in parallel looking for repeating patterns. Looking for integration related terms, a new layer of codes were applied: “Technical integration” was mapped to component integration, testing, interface and also to legacy (and other) systems. “Stakeholder integration” was mapped to stakeholders, actors, users, customers, organizations and vendors. The integration context was also coded; whether it was internal (to project) or external (to business process and information infrastructure) integration.

Table 3. Data collection

<table>
<thead>
<tr>
<th>Phase/dates</th>
<th>Activities</th>
<th>Stakeholders</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td>Initial meeting with management to agree on objectives and procedures in the study.</td>
<td>Line manager</td>
<td>Project objectives and plans</td>
</tr>
<tr>
<td>Sept. 2001</td>
<td></td>
<td>Project managers</td>
<td></td>
</tr>
<tr>
<td>Phase 2</td>
<td>Workshop with project and business stakeholders to get the broad picture, followed by separate interviews</td>
<td>Project manager</td>
<td>Status reports</td>
</tr>
<tr>
<td>Nov./Dec. 2001</td>
<td></td>
<td>Project group</td>
<td>Technical and architecture documents</td>
</tr>
<tr>
<td>Phase 3</td>
<td>Separate interviews with stakeholders to construct full timeline in project</td>
<td>Project manager</td>
<td>Status reports</td>
</tr>
<tr>
<td>Sept./Oct. 2002</td>
<td></td>
<td>Project group</td>
<td>Project evaluation report</td>
</tr>
<tr>
<td>Phase 4</td>
<td>Last round of interviews. Validation meeting to confirm and discuss findings.</td>
<td>Line manager</td>
<td>Case description</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Business users</td>
<td></td>
</tr>
</tbody>
</table>
Validating the Case Description

Suggested mechanisms and patterns were validated using criteria suggested by Klein and Myers (1999).

The principle of dialogical reasoning addresses the validity problem of the relationship between the assumptions behind the research design and the actual findings. In the case study, this aspect was important in the coding process described in the previous section. Moving from codes to patterns, systematic attempts were made to construct the opposite pattern and ask whether it could make sense on the same data.

The principle of multiple interpretations calls for sensitivity in interpreting conflicting accounts of the same case. In the case description, special care was taken to analyze instances where data was contradictory; for example, when developers and user representatives had different accounts of what had happened.

The principle of member verification (Ngwenyama, 1998) was implemented as follows: After the initial workshop with important stakeholders, a timeline for the project was constructed. This was sent to participants for verification. After about a year of studying the project, a preliminary case description was written and commented on by stakeholders. Last was a validation meeting, where project managers, developers and business users were invited to assess and comment on the final case description.

The SAS Content Management Interface (CMI) Project

The case is described from a socio-technical integration perspective, focusing on the stakeholder and technical integration challenges.

SAS is an international airline carrier based in Scandinavia. It is owned by the Swedish, Danish and Norwegian governments (50%) and private investors. In 2003, the SAS group had 34,000 employees, and turnover was 57 billion Swedish krona (SEK).

As a part of the marketing division, SAS established in 1999 the Electronic Direct Channel (EDC), directed at selling airline tickets and hotel reservations on the Internet. The EDC owns the SAS home page, including campaigns and online booking services. Simplified, the e-business process consists of these steps: Marketing air tickets in different national markets on the scandinavian.net Web site, and receiving electronic orders. Feedback mechanisms are Web hits on the home page and actual bookings.

The first generation of the SAS Web site was primarily available in Scandinavia, with content managed by editors using HyperText Markup Language (HTML) and FrontPage as publishing tools. In 2000, acknowledging the commercial potential of Web-based booking, SAS decided to establish a Web-based marketing channel in all important SAS markets, including Europe, Asia and the Americas. To support this new business process, a new content management and publishing solution was needed, and implementation was based on the Vignette platform. A project was initiated, with the following objectives:

- To establish a Web-based marketing channel in all important SAS markets, including Europe, Asia and the Americas.
- To enable the editors to publish materials and campaigns with an easy tool, a CMI.
- To integrate this new system with SAS’ (legacy) booking system and the European Amadeus booking system.

In spring 2001, the IS development project was set up by SAS’ daughter IT company, Scandinavian Information Group (SIG), with a project group of five: one project manager, one Web designer and three programmers. In parallel, an SAS project group, consisting mainly of marketing editors, was established to provide a strong business and user link with the development group.
Following a corporate decision from 1999, the project was planned within the RUP framework. Each iteration was set up to follow the disciplines in RUP, starting with a revision of requirements; proceeding with design, coding and testing; and ending with an increment, a temporary release, to be validated by users. Also, in accordance with RUP, a stakeholder and risk analysis was done.

### The First Two Iterations: Unsuccessful Stakeholder Integration

The initial requirements specification, written by IT personnel at SIG, contained 14 use cases for the CMI application. In two workshops, the two project groups extended the number of use cases into 20 detailed ones. Then they started working on a graphical prototype to visualize the use cases.

The workshops were not very successful, and they were interpreted differently by the actors: The SAS project manager, who was now elaborating the requirements specification, was moderately satisfied:

*The workshop in the first iteration was OK because it gave the users an impression of the system. The workshop in the second was useful, but we were not able to show the users how the system would work.*

Some of the editors felt alien to the whole concept:

*We spoke different languages, and they had no idea how we worked. We were polite, and there was no conflict, but that was how we felt. We thought we might get it straight later on in the process. The use cases were very – theoretical, and it was hard to participate. Also, I was so busy running the existing site that I had no time to spend on this, really. I felt guilty because of that, because I should have given much more input.*

Another editor said:

*Use cases focused on the new system, not on how things were solved today. Development was system oriented, not on the work process.*

The developers later said:

*Of course, the graphical prototype should have been a full architectural prototype – but this was not possible, because the presentation engine component was not ready. In addition, the editors did not really prioritize the workshops.*

Not surprisingly, the results were unsatisfying. Nobody felt that the graphical prototype was useful. The presentation engine component, which was crucial for creating a page, was developed in another project in Copenhagen, and it was seriously delayed. Thus, by the end of the elaboration phase, the two main goals were not reached: The users and developers did not have a unified view of the system, and the architecture of the system was not stable. The situation was analyzed and the risks were assessed by the customer and the project managers. Because of time pressures, they decided to proceed, and took care to prepare as much as possible to integrate the missing component later.

### Iterations Three and Four: Successful Internal Technical Integration

No code had been produced in the first two iterations. In the third iteration, the project group got a better grip on the technology, and started to work more closely with the SAS project manager, who was now sitting in the same room. This iteration produced the basic CMI functionality, enabling the CMI users to upload content to the content database.

In the fourth iteration, the first release of the presentation component arrived, and the crucial
functionality for creating pages was developed. In a few intense and informal work sessions, a design was developed as the application was prototyped. Commented one of the developers:

*When the SAS project manager really joined the team, the whole atmosphere changed. We were able to experiment with screens and solutions at a practical level. Also, it was important that he really understood the technical difficulties involved. We were sitting long hours together, solving real problems. It was very productive and also great fun!*

Test cases were also produced, and at this time the project group felt things were on the right track. While the user group was not involved, the SAS project manager communicated frequently with the editors, assuring them that the project was on track.

After the start of the fourth iteration, the September 11, 2001 terrorist attacks in the United States (U.S.) shook the whole airline industry, and as a result of this, combined with a tight project deadline, the project was scaled down. Only 10 of the initial 20 use cases would be developed. By the end of November, the user group started testing the system. Towards the end of the iteration, the SIG project manager went on maternity leave, and the SAS project manager was made in charge of the whole project.

**Iteration 5: Improvised Integration**

So far, the project had proceeded more or less as planned. But now a different phase started, which was not characterized by planned iterations but rather by problem solving and improvisations. The project manager commented:

*At a certain point in a project the focus changes: Budget and plans are no longer the main consideration, and everything is focused on getting the product finished.*

There were two reasons for this. First, the project experienced external technical integration problems. In the fourth iteration, the project waited for the completion of two important software components that were programmed by another SIG project in Copenhagen. Both were seriously delayed and not stable until spring 2002. There were also other technical problems. SIG had four different technical environments: Development, integration testing, production testing and production. In this case, the test environments did not fully support the CMI, and intermediary solutions had to be made. Complicating matters even more, a memory leak occurred, which took a long time to identify and fix. These problems made testing more difficult. Second, after beta testing had started, there were a rising number of change requests from the user group, and changes were introduced frequently during spring 2002. One of the editors said:

*From the end of February, we started testing at our workplace. At that time I had no mental picture of the whole systems, only fragments. In March, when we started to load real data, I was able to understand the workings of the system. Until then, it was like looking at the system through a pair of binoculars.*

As the editors got more experience, the number of change requests increased. Said the Oslo editors:

*The period around the start of the system we sat together with the developers and really understood the whole system. This was great, and the first time we really could give real input.*

In March, there was a 2-day course in Stockholm for all the editors, at that time numbering around 30. Most of them were introduced to the system, and in spite of technical stability problems and long response time due to slow interfaces (APIs) in the Vignette platform, the course was perceived by the EDC and CMI project as rather
successful for most of the editors. A few of them were less motivated, and lacked the basic IT user skills. After the course, the editors went home and started to load materials into a test database, which was later set into production. In this period, the CMI project worked hard with error corrections and use case change orders. The project manager said:

*Many new features were wanted from editors, both Scandinavian and the others, especially navigation features tightly connected to their work processes, page search and design. We were surprised by the volume of change orders.*

**Results**

Everything was set into production May 30, 2002, and has been in stable production since. The three goals of the project were achieved: a Web-based marketing channel in all SAS markets was established; the marketing editors in this organization had access to an easy-to-use tool to tailor the marketing and selling of SAS services to their local markets; and the Web pages were integrated with the booking systems, enabling customers to buy tickets online. However, during the first 6 months of production, there were still a large number of change requests from the marketing editors.

The case study highlights a number of managerial challenges associated with integration. These will be discussed in the next section.

**DISCUSSION**

This section first discusses the findings of the case study, in terms of managerial challenges of integration. Then some implications for practice and research are proposed. Finally, some limitations of the findings are assessed.

**Table 4. The managerial challenges of integration**

<table>
<thead>
<tr>
<th>Integration type</th>
<th>Managerial challenge</th>
<th>Support in RUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal technical integration</td>
<td>Step-wise internal integration of components needs careful planning and competent execution.</td>
<td><em>This is well supported in RUP and in traditional project management techniques.</em></td>
</tr>
<tr>
<td>Internal stakeholder integration</td>
<td>Step-wise integration of project team members calls for sensitivity, especially in relation to user representatives.</td>
<td><em>This is partly supported in RUP.</em></td>
</tr>
<tr>
<td>External technical integration</td>
<td>Integrating to the technical infrastructure is outside the direct control of the project manager. It calls for careful planning and risk management, but also for improvisations.</td>
<td><em>RUP gives some high-level support, but not at the operational level.</em></td>
</tr>
<tr>
<td>External stakeholder integration</td>
<td>Integrating to the business process is also outside the direct control of the project manager. It calls for planning and risk management, but also for political skills and improvisations.</td>
<td><em>RUP gives some high-level support, but not at the operational level.</em></td>
</tr>
</tbody>
</table>
Managerial Challenges of Integration

The findings of the case study are discussed within the framework of the four integration types (presented in Table 1). They will also be related to the analysis of integration support in RUP presented in an earlier section.

In the described case, integration is seen as a process. It was partly or fully achieved through different managerial interventions, supported by the iterative and incremental structure of RUP. The managerial challenges associated with the four integration types are discussed below and summarized in Table 4.

Internal Technical Integration

This aspect was, after some initial problems, solved very nicely in iterations three and four. The CMI system was built incrementally and controlled, utilizing the rich mechanisms in RUP for step-wise integration and testing.

Findings: For the technically competent project manager, a software engineering framework such as RUP provides sufficient support for step-wise and controlled technical integration.

Internal Stakeholder Integration

At the start of the CMI project, the most central users in the business process, the Scandinavian editors, were drawn into the project and given a central role in the workshops. The workshops in iterations 1 and 2 were not successful, mainly because the project managers failed to convince the editors of the need for a new system. The project group responded to this with a certain degree of isolation, establishing a close team with the SAS project manager, who “changed sides”; that is, who identified more with the project than with the business process.

The progress of software development benefited very much from this cooperation. However, when the number of change requests increased in the later parts of the project, it was an indication that the iterative approach was not working optimally regarding user requirements.

Findings: Iterative IS development projects are not based on a frozen requirements specification, but rely on a step-wise interaction with business representatives through the iterations. Thus, it is critical to align project goals with business and user goals even after the initial top-down analysis. RUP does not give sufficient support for this task, as shown earlier.

External Technical Integration

The CMI project suffered from being dependent on components that were delayed. Thus, while skillfully planned, the step-wise technical integration was not working properly. Instead, the external integration problems appeared at a time when the editors tried to integrate the system into their organization, and it threatened the whole implementation.

The main technical problems arose from the fact that the development environment, and also the test environment, was configured differently from the production environment. In the test phase, this led to a series of small but important problems of instability. In the production phase, it led to the cache problem. As the project was designed, this kind of risk was hard to avoid, because a test environment mirroring a changing production environment would have been exceedingly expensive. In the end, the problem was gradually solved by a competent data center, using problem-solving techniques, and not a structured development process.

Findings: Integrating with a large, existing information infrastructure calls for careful planning, but also for improvisations. Often, the infrastructure is subject to change during the project, and it is therefore impossible to integrate in a step-wise manner. Obviously, the project man-
ager has to comply and adjust to these changes, whether or not they were planned. This aspect is underrated in RUP.

**External Stakeholder Integration**

The main target group for the CMI specification was the international editors, who needed a standardized and simple interface. Unfortunately, during the first four iterations they were not yet appointed, and could not participate. Thus, the continuous change of the business process made step-wise stakeholder integration unfeasible.

On the other hand, the international stakeholders were successfully enrolled in the project from March 2002, and during the busy, and partly improvised, period towards the production date, most of their requirements were satisfied. It should be noted, however, that during the first 6 months of production, there were still a large number of change requests, indicating that important business needs were not covered in the specification. It should also be noted that the late integration of stakeholders was done by improvisation.

*Findings:* Integrating with the business process calls for political skills and improvisations. Political skills are needed to build the alliances that make the solution work. Improvisations are necessary to integrate with an unstable process.

Of course, it is unreasonable to blame RUP for the late integration problems in the CMI project. But the “extended iteration” in the project, where the structured method gave little support and the project manager resorted to improvisation and problem solving, also indicates that this challenge is underrated.

*Summing up:* IS development in a complex and unstable technical and business environment, as illustrated in the SAS case, highlights the need for step-wise socio-technical integration.

Iterative IS development has an interesting potential to support this challenge, because the iterative structure allows for step-wise learning and incremental improvement. It also, to a certain degree, acknowledges the emergent nature of socio-technical interaction. However, in its present condition, RUP is too heavily influenced by its purely technical inheritance to give sufficient support to the full range of iteration challenges.

**Implications for Practice**

The CMI case highlights an important insight for practitioners: The integration challenge of iterative software development projects is larger and more complex than is described in the research and textbook literature. Even the very competent CMI team experienced serious problems with external integration. For practitioners, this is an important factor in risk management: To succeed, the project manager must manage something that is basically outside his or her control.

The root of this risk is the fact that both the business process and information infrastructure are unstable in most enterprises. As illustrated in the CMI case, the focus in a development project may change considerably over time, making it hard to plan in detail. And changes in the business process and information infrastructure should be treated as a normal feature, not as an unexpected incident. For the project manager of iterative projects, this aspect also presents opportunities, but in a world of tight schedules and budgets, the external technical and stakeholder integration is a significant risk.

For the IS project manager, this implies that risk management—focusing particularly on external integration—is as important as internal planning and control.

**Implications for Research**

At a general level, these risks are acknowledged in the IS development research (Avison & Fitzgerald, 2003) and in iterative software engineering frameworks (Jacobson et al., 1999; Stapleton, 2003).

However, as illustrated in the SAS case, these risks are not appendices in handling a
primarily technical task; on the contrary, they constitute a significant threat to the success of such projects. As the RUP analysis in the earlier section shows, the support in RUP is not sufficient to handle this challenge, in spite of some minor contributions (e.g., Kruchten, 2003).

Thus, on a practical level, RUP needs stronger support for external stakeholder and technical integration; in particular, in the construction and transition phases. In practice, this means emphasizing the external component/legacy system issues in the architecture description, and supplementing risk lists, iteration plans and task descriptions with activities that address business process stakeholder issues more explicitly. At the same time, it should be acknowledged in the RUP guidelines that the external integration cannot be planned in detail.

Planning and Improvising

External technical and stakeholder integration were seen to depend on both planning and improvisations to succeed. Finding that improvisations are important in change processes is congruent with the emergent view on organizations and with other related IS research (Ciborra, 1997, 2000; Ngwenyama, 1998; Orlikowski, 1996). However, it should not lead us to conclude that better planning is unnecessary or impossible. As the CMI case shows, improvisation puts a lot of pressure on a project manager, and introduces more risk into the project. The strength of an iterative software engineering framework is to reduce that risk, through step-wise integration. Thus, it is important to extend the capabilities of RUP and other frameworks, to include better external integration.

Limitations

Limitations to the findings presented in this paper derive from the research approach. LPR is aimed at developing contextualized theory (Pettigrew, 1985). This suggests that the area of validity is the context of a large organization with an iterative IS development process in a setting of unstable business processes and a complex technical infrastructure.

CONCLUSION

The objective of this study was to understand the integration challenge in modern iterative IS development projects. Two questions were asked:

1. How can the IS project manager integrate the new information system with the business process and the information infrastructure?
2. What integration support is there in the software engineering frameworks, like RUP?

Integration was defined in a socio-technical perspective, and four types were suggested: external and internal stakeholder integration, and internal and external technical integration. A longitudinal case study of an IS development project in the airline industry identified and highlighted some managerial challenges of integration. Findings include:

- Internal technical and (to some extent) internal stakeholder integration is well supported by traditional project management techniques and software engineering frameworks such as RUP. The iterative approach enables a step-wise and controlled internal socio-technical integration.
- However, the challenges of external stakeholder and technical integration is underrated in IS development research. Integrating a new system with an unstable business process and a large, existing information infrastructure calls for careful planning, but
also for improvisations, since it is outside
the direct control of the IS project manager.
This is not well supported in RUP.

These challenges are suggested to constitute a
significant risk for the success of modern iterative
IS development projects. For IS project manag-
ers, the implications are that risk management
of external integration is an important part of
iterative project management. For software en-
gineering research—in particular for RUP—the
findings indicate a strong need to provide stronger
support for external stakeholder and technical
integration; in particular, in the construction and transition phases.

Further research could study other iterative
projects and look closer into which integration patterns emerge. One might also investigate new
models for how IS project managers may contrib-
ute more actively to align the development projects and the business innovation, as recently done by
Boehm by the introduction of value-based soft-
ware engineering (Boehm, 2003). This is an area
well suited for cooperation between researchers and practitioners.

ACKNOWLEDGMENTS

I am indebted to the different stakeholders of the
SAS project, in particular Christian Mejlvang,
Jorun Akerberg and Gunilla Johansson, for ac-
cess to the project and for commenting on vari-
ous versions of the case description. Further, the
comments from the anonymous reviewers of the
International Journal of Technology and Human
Interaction were very helpful. Thanks also to the
following scholars and practitioners for comments:
Peter Axel Nielsen, Bjørn Erik Munkvold, Hen-
nning Denstad and Cecilia Haskins.

NOTE

An earlier and shorter version of this paper was
presented at the 37th Annual Hawaii International
Conference on System Sciences, January 5-8,
2004, Big Island, Hawaii.

REFERENCES

systems development: Methodologies, techniques

Beck, K. (2000). Extreme programming explained:
Embrace change. Boston: Addison-Wesley.

Blomberg, J. (2001). Using the RUP for enterprise
e-business transformation. The Rational Edge.
com/developerworks/rational/library/content/
RationalEdge/jan01/UsingtheRUPforEnterpri-
seebusinessTransformationJan01.pdf

development and enhancement. IEEE Computer,
21(5), 61-72.

Boehm, B. W. (2003). Value-based software en-
gineering. ACM SIGSOFT Software Engineering
Notes, 28(2), 1-12.

the concept of strategic alignment. Scandinavian

Oxford University Press.

Coakes, E., & Elliman, T. (1999). Focus issue on
legacy information systems and business process
engineering: The role of stakeholders in managing
change. Communications of the AIS, 2(4).

Boston: Ernst & Young.


*This work was previously published in the International Journal of Technology and Human Interaction, edited by B. C. Stahl, Volume 2, Issue 4, pp. 1-16, copyright 2006 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 6.8
Human Factors for Networked and Virtual Organizations

Vincent E. Lasnik
Independent Knowledge Architect, USA

“There are simple answers to all complex problems—and they are uniformly wrong.”

INTRODUCTION

One of the central problems and corresponding challenges facing the multi-disciplinary field of networked and virtual organizations has been in the construction of theory-grounded, research-based taxonomies for prescribing what particular strategies and approaches should be employed when, how, and in what combination to be most effective and efficient for specific business domains, organizational structures, and enterprise-wide performance objectives.

An Imperative Requirement for Human-Centered Design

Our workplaces—and certainly our daily lives—have become increasingly cluttered by a vast array of ubiquitous “high technology” products from CD/DVD/M3P players, personal data assistants (PDAs), laptops, and recently mini-palm-top computers, cell phones, digital cameras, graphic video games, voice mail, e-mail, and Web-accessible portable devices, electronic car keys, answering machines, microwaves, alarm clocks, printers, fax machines, pagers, programmable robotic vacuum cleaners, and multiple-component home entertainment systems—to name only a few. Learning to use many of these devices, far from being transparently easy, requires a steep learning curve punctuated by psycho-emotional intimidation, trial and error, frustration, and a personal sense of belittlement in the face of the impersonal, ubiquitous, and insurmountable dominance of machines. An impactful January 2001 cover story in U.S. News & World Report—“Overwhelmed by Tech!”—called these proliferating devices “fiendish new instruments of mental torture” and said that “consumers may not expect all their new gadgets and gizmos to be fun, but they are demanding that at least they don’t make them feel like idiots” (Lardner, LaGesse, Rae-Dupree, & Roane, 2001). This is particularly salient for members of networked and virtual organizations.
who rely so critically upon their tools, systems, and user interfaces to be creative and innovative knowledge workers.

It might be an elusive goal to seek a single standard or set of guidelines that perfectly integrates all of the different descriptive practice (i.e., case scenarios/situations) and prescriptive management techniques and strategies within an easy-to-implement solution paradigm for every creative, innovative, and knowledge-based requirement. Up to now, it has been a nearly intractable challenge to establish unequivocally-accepted technical standards across ANY industry, as current competing specifications for HD-DVD, HDTV, widescreen format aspect ratios, Web-streaming formats, SCORM compliance, technology transfer and transportability, product interoperability, and reusability continue to persist. While it can be argued that over time many competing technology systems find some pragmatic, functional accommodation to co-exist within the same universal marketplace (e.g., AC and DC, AM and FM, Apple and Microsoft), idiosyncratic conceptual paradigms like management theories and organizational models have ultimately “agreed to disagree” in addressing the same, similar, or related issues and problems from different perspectives in different enterprises and nations. And this is not only intelligent and mature but probably wise since the compelling task of creating comprehensive explanatory models of complex phenomena that are parsimonious and robust without being overly reductionist—and concise and cogent without being over-simplifications—continues to be formidable.

**HUMAN FACTOR DESIGN PRINCIPLES: AN OVERVIEW**

On November 3, 2005, the first World Usability Day was held by the Usability Professionals’ Association to promote “user-centered design and every user’s responsibility to ask for things that work better.” Many permeating technologies within networked and virtual organizations share the amberification paradox of textbooks, newspapers, Whitepage listings, and nightly television news programs: they are all out of date as soon as they are produced and become like artifacts hermetically sealed in amber and suspended in time from the day of their encasement. Their “one size fits all” designs are neither adaptive nor adaptable and promulgate a top-down technological imperialism penetrating every aspect of their use. The communication, collaboration, and interaction infrastructure of organizations thus remain sorely challenged with enduring problems of usability, learnability, accessibility, and adaptability. As the function and form of products undergoes increasingly rigorous scrutiny, one important design goal is emerging as a paramount priority: improving the usability of products, tools, and systems for all stakeholders across the enterprise. From the perspective of the new science of human factors in user experience design (UXD)—lowering the cognitive workload by simplifying the system interface will contribute to a less stressful, more confident task performance, and a more effective and efficient user experience.

**Enhancing Information and Communication Design Efficiency**

In user-centered design, the conceptual model of the user becomes the superordinate principle guiding the design process and outcomes (Kuniavsky, 2003). Usage-centered design is different: it focuses primarily on the empirical (i.e., observable and measurable) goal-based, task-driven behavior of users and the structuring activities, procedures, processes, and corresponding information architecture required to optimize the effectiveness of the user to efficiently accomplish those functional goals. By integrating the human factors engineering “systems approaches”, it may be possible to optimize the beneficial design quality of products and services from the perspective of the user’s
experience (Constantine, 2001; Constantine & Lockwood, 1999; Wickens et al., 2004).

While examining ways to improve the signage, self-directed wayfinding, and user-centered design of interactive public information environments like museums, science centers, and shopping malls, C. G. Screven (1999) synthesized a number of highly pliant and insightful principles that could be effectively repurposed for the home-based infosphere we are discussing. Screven’s core UXD metric is design efficiency—described as “the average time it takes viewers to find and process message-related information” (p. 147). Efficient designs (a) conserve the time and both mental (i.e., working memory) and physical effort needed by users to find, access/retrieve, quickly understand, and respond to messages and content, in effect focusing their attention and augmenting their real-time decision-making capacity; (b) reduce fatigue and “maturation effects” by providing affordances with low cognitive loads to quickly engage and involve the user in self-directed behavior that enhances their interactivity with the exhibit’s information (i.e., less noise, more signal); and (c) generally improve all aspects of the information designs themselves by systematically employing larger, easy-to-read fonts, less dense and less difficult textual content (e.g., using familiar, jargon-free vocabulary), good contrast and consistent layouts (i.e., with good “chunking”, white space, use of colors and clear mnemonic cues, and reduced visual clutter), and precise, unambiguous headings and labels. An optimal high design efficiency is thus characterized by (a) a high ratio of needed and salient information to unneeded, superfluous information stimuli to ease working memory load and reduce affective stressors, and (b) facilitates users’ abilities to process and absorb new information (Preece, Rogers, & Sharp, 2002; Ware, 2004).

Screven is an advocate for thoughtfully designing human-centered information spaces that reward and reinforce the fun and enjoyment of goal-directed experiences while intelligently reducing the aversive stimuli associated with the disconnects of failure, confusion, frustration, avoidance, task-time and effort, media overload, product complexity, and the depersonalizing emotions engendered by being continuously fettered by seemingly ineluctable, inefficient information design solutions. The enactive, intentional, unifying higher-order problem-solving endeavor is design itself—and numerous universal principles, exemplars, and epitomes of design have emerged (Lidwell, Holden, & Butler, 2003). “Designing is, therefore, more than ordering and arranging, more than constructing. It is composing. It is using the codes and pattern languages of a domain to create wholes with not only parts and relationships but also ordering-underlying principles” (Rowland, 2004, p. 40). Critical in this human-centered design process are systems thinking, creativity and evaluative judgment, metacognitive awareness, and the seemingly paradoxical nurturance for an eclectic, broad-minded tolerance for ambiguity (ACM, 2005).

Respecting The User Experience

In their insightful book “Designing From Both Sides of the Screen: Don’t Impose—Respect Mental Effort”, Isaacs and Walendowski (2001) analyzed screen-based user interfaces. They noted that every user choice, every click increases working memory load, and that every separate button increments functional overchoice and visual clutter. There is also an emotional, affective stress factor that burdens users caught in overly complex interfaces (i.e., “If they keep bumping their heads against the same problem over and over again, their tolerance capital declines each time”), and this is particularly apparent in networked and virtual organizations that depend upon sophisticated, interactive synchronous and asynchronous IT systems and tools. User experience designers need to be sharply focused on developing devices and systems that reduce (not increase) working memory loads, eliminate unnecessary features,
Human Factors for Networked and Virtual Organizations

and buttress the confidence and competence of basic user task behavior (Stone, Jarrett, Woodroffe, & Minocha, 2005). Designers should leave their product users "as much mental energy as possible so they can flow with their task and forget about the technology entirely." Their summary design guidelines recommend that designers: (a) reduce wayfinding memory by using visual elements like buttons, menus, icons, and legends sparingly; (b) make common tasks clearly visible and easy to find while hiding infrequent tasks (i.e., the main screens and remote control interfaces should “offer people quick access to the things most people do most often, while giving newcomers an easy way to get started”); (c) give users feedback (e.g., combine audible and visual cues simultaneously) and display signs of task progress, build-in good “error recovery” so users can “undo” an operation; and finally (d) if a command cannot be carried out quickly, allow users to interrupt the process and return to a previous stable system state. These design approaches are corroborated by others in the field (Kalbach, 2004; Shneiderman, 2004) who are attempting to put more “fun” into product functionality. It would also be reasonable to argue that such improved systems would also augment the sense of stakeholders’ competence, emotional well-being, and overall enjoyment of their work environments—further adding an affective resource to enhance their task performance. In describing the recent interest in usability hedonics, Hassenzahl, Beu, and Burmester (2001) declared that “the most basic reason for considering joy of use is the humanistic view that enjoyment is fundamental to life.”

DEVELOPING A HUMAN FACTORS TAXONOMY

The taxonomic concept derives from the Greek words taxis (for arrangement, order) and nomos (law). Every serious taxonomy is an organizational scheme that includes a system representing structure, order, and relationship. Some form of hierarchical structure is generally defined, but this may be multi-dimensional and nonlinear in form. The purpose, domain, attributes, and granularity of schema vary but all taxonomies attempt to provide a robust (i.e., logical, coherent, cohesive, internally-consistent) architecture. Prominent examples include the widely-adopted schema of Carl Linnaeus (biology) and Dmitrii Mendeleeev (The Periodic Table of Elements). Most taxonomies contain their own nomenclature for describing the taxons (singular) and taxa (plural) that correspond to formal units in the classification scheme, such as kingdom, phylum, class, order, family, genus, species (adapted from Linnaeus). Taxonomies may evolve over time. Neither the systems of Linnaeus nor Mendeleev are exactly in their original form when they were first presented, but they are fundamentally and substantially the same in all relevant aspects and overall structure, changing only as our knowledge of science changed over time to add additional granularity to the taxons and taxa of their brilliantly original and enduring descriptive taxonomies.

Our emerging organizational taxonomies (e.g., such as those referenced in Table 7.4 in Thamhain, 2005, pp. 180-185) will need to integrate the many complex factors involved in the human-product-project triad, including the dynamic human factors issues of manager/employee, the special nature of the particular product/service/industry domain, and the implicit assumptions of the successful networked and virtual project team environment. To illustrate a simplified taxonomic analogy of the basic conceptual issues (using seven common architectural concepts as points of comparison), Table 1 is presented.

Principles of Universal UXD

In 1997, The Center for Universal Design at North Carolina State University produced a comprehensive set of guidelines to improve the “design of products and environments to be usable by all
people, to the greatest extent possible, without the need for adaptation or specialized design.” Funded by public federal grants from The National Institute on Disability and Rehabilitation Research and the United States Department of Education, a working group of architects, product designers, engineers, and environmental design researchers collaborated to create seven core principles to help guide a wide range of design disciplines including environmental, industrial, and communications design. These user-centered guidelines were intended to be systematically applied to evaluate existing product designs, to re-focus the future design process of new products, and to educate both designers and consumers about the characteristics of more usable products and environments. They are particularly appropriate for application to all networked and virtual organizations.

One version of these Principles of Universal (User Experience) Design are presented next with the name of the principle and a concise definition of the key concept embodied in the principle (Center for Universal Design, 1997). While most of these core guidelines should be applicable to nearly all design problems, a reasonable caveat is that not every sub-component may be relevant to every design solution.

**Principle One: Equitable Use**
- The design is useful and marketable to people with diverse abilities. Provide the same means of use and provisions for privacy, security, and safety to all users.

**Principle Two: Flexibility in Use**
- The design accommodates a wide range of individual preferences and abilities. Provide choice in methods of use, facilitate the user’s accuracy and precision, and provide adaptability to the user’s pace.

---

**Table 1. Simple analogical taxonomy of post-modern organizational components**

<table>
<thead>
<tr>
<th>BUILDING ARCHITECTURE</th>
<th>ORGANIZATIONAL ARCHITECTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buildings, structures, bridges: Core functional blocks of environment</td>
<td>Institutional framework, functional system, operational areas (3 layers)</td>
</tr>
<tr>
<td>Macro-design form, leitmotif, higher-level conceptual treatment, style</td>
<td>Organizational design models (e.g., Hierarchical, Matrixized, Projectized)</td>
</tr>
<tr>
<td>Purpose &amp; function of building</td>
<td>Project evaluation and selection; Core product/service focus of organization</td>
</tr>
<tr>
<td>Physical properties of materials</td>
<td>Properties of core technologies used to design/develop organization’s products and services</td>
</tr>
<tr>
<td>Patterns of interior/exterior space</td>
<td>The work environment, both physical &amp; psychological; interaction patterns</td>
</tr>
<tr>
<td>Structural lifecycle (repair, replace, repurpose, modify, adapt, change)</td>
<td>Iterative project/product design and improvement; cross-functional concurrent engineering &amp; integrated product development approaches</td>
</tr>
<tr>
<td>Settlements, zones, cities, urban, conurban, &amp; suburban, rural districts</td>
<td>Systems, modules, units, teams: Underlying structures of projects and intra-enterprise inter-relationships</td>
</tr>
<tr>
<td>Power, water, air, transportation, communication, food, &amp; waste mgmt</td>
<td>Capital, material, and human resource management infrastructure</td>
</tr>
</tbody>
</table>
Principle Three: Simple and Intuitive Use
• Use of the design is easy to understand, regardless of the user’s experience, knowledge, language skills, or current concentration level. Eliminate unnecessary complexity, be consistent with user expectations and intuition, accommodate a wide range of literacy and language skills, arrange information consistent with its importance, and provide effective prompting and feedback during and after task completion.

Principle Four: Perceptible Information
• The design communicates necessary information effectively to the user, regardless of ambient conditions or the user’s sensory abilities. Use different modes (pictorial, verbal, tactile) for redundant presentation of essential information; differentiate elements and maximize “legibility” of essential information; provide compatibility with devices used by people with sensory limitations or disabilities.

Principle Five: Tolerance for Error
• The design minimizes hazards and the adverse consequences of accidental or unintended actions. Arrange elements to minimize hazards and errors: most used elements, most accessible; hazardous elements eliminated, isolated, or shielded; provide fail safe features.

Principle Six: Low Physical Effort
• The design can be used efficiently and comfortably and with a minimum of fatigue. Minimize repetitive actions and physical effort.

Principle Seven: Size and Space for Approach and Use
• Appropriate size and space is provided for approach, reach, manipulation, and use regardless of user’s body size, posture, or mobility. Provide a clear line of sight to important elements for any seated or standing user and adequate space for the use of assistive devices.

While it is true that the practice of design involves more than consideration for the usability requirement per se, a judicious inclusion of these guidelines to the iterative information-interactivity-media design process of communication technologies for networked and virtual organizations would ostensibly accommodate the needs of as many users as possible and in effect meet or exceed many of the collateral concerns about the economic, engineering, cultural, and environmental impact (i.e., efficiency) of the design solution. Moreover, while these universal principles are not exhaustive in scope or detail, they provide a helpful starting point for a heuristic evaluation of any human-system product’s user experience portfolio.

Organizations Without Human Factors Environments

Prior to the advent of networked and virtual organizations, the very strengths of large “Industrial Age” functional organizations were consistent with their weaknesses (e.g., risk-averse, poor cross-division communication and cooperation, limited technology transfer, limited flexibility to respond to rapid change trends, fettered high-level decision-making capability, and generally inflexible “time-in-grade” hierarchies of power and influence). In dysfunctional enterprises, the rate of change is deliberately slowed to that of a moving glacier, and the future tends more often than not to look much like the past because without the implementation of human factors engineering their strategies repeat the past, aggrandize the past, recapitulate the past, preserve the past, reproduce the past, and ultimately ensure that the future and the past appear indistinguishable.
Today, “Information Age” and emerging “Knowledge Age” environments clearly require a new breed of employee and management. Individuals on all organizational levels must be much better educated, more highly skilled, predisposed to expect constant change, and be more committed to the underlying success of the broader enterprise. Specifically, the managers in these technology-intensive endeavors must be able to simultaneously: (a) direct immensely sophisticated, increasingly complex internal processes and resource-sharing operational structures (e.g., projectized and matrixized), and at the same time (b) motivate, engage, and flexibly integrate their intelligent, creative, and individual career-focused workers into cohesive, collaborative team wholes that can mutually accomplish high-demand goals under their ever-tightening production schedules and technical requirements. They must be superb leaders of both “men and machines”.

The Need to Innovate

Managers of networked and virtual organizations understand the need to innovate but often struggle to pragmatically support creative processes within demanding, requirements-driven business environments. Information architects, knowledge developers, media designers, and usability specialists often face similarly demanding contexts—but must resourcefully find ways to support conceptually creative projects and innovative processes reliably and successfully (Hughes, 2002). These new "pliant" processes must either be invented or repurposed (i.e., via breakthrough human factors engineering scaffolds), but in every case one or more persons are involved at the core (Henderson & Harris, 2000). In fact, a new professional role is emerging that brings knowledge construction and production to the epicenter of information designers and managers (Morrogh, 2003). Ostensibly then, managers can learn from designers about the process of innovation and as the need for this “learning organization” grows—economic activity moves into the realm of knowledge work, where materials are intangible and human innovation and idea-generation are the primary source of value-added creation. This is, after all, what the promise of virtual and networked organizations is able to achieve.

Principles of Corporate Creativity for the Knowledge Age

As described by Robinson and Stern (1997), the basic constructs of fostering organizational creativity include: alignment, self-initiated activity, unofficial activity, serendipity, diverse stimuli, and within-company communication. Like the principles for universal product design, these human factors principles can ostensibly augment creativity and productivity in networked and virtual organizations. This set of design considerations includes:

Principle One: Alignment

• An organization’s performance depends on its alignment—the degree to which the interests and actions of every member mutually support’s key goals. The organization cannot be consistently creative unless it is strongly aligned. Moreover, the most critical step in aligning an organization is recognizing the value of alignment and that it has to be done.

Principle Two: Self-Initiated Activity

• People have a natural desire to be creative so most unexpected creative acts come from self-initiated activity. The key to organization-wide self-initiated activity is an effective system for responding to member’s intrinsically-motivated ideas. Networked and virtual organizations can become highly efficient in this regard to the degree that the formal hierarchy becomes flexible, flatter,
Pliant and more open to evoke, foster, facilitate, encourage, accept, and reward fresh ideas, original initiatives, and “value added” innovations for processes and products.

**Principle Three: Unofficial Activity**
- Within virtual and networked organizations, many innovative and creative ideas and approaches begin with unofficial, informal, and independent human activity during which an idea is worked on without direct or official support. Unofficial “under the radar” status reduces disincentives to ongoing creative communication associated with stifling micro-management and inhibitory gatekeeping, rigid reporting hierarchies (i.e., that strictly follow the org chart), peer surveillance, political turfism, performance pressures, and personal jealousies. Unofficial activity can provide a safe haven for generating and incubating ideas, and leads to better decision making about which new high-tech project initiatives to support.

**Principle Four: Serendipity**
- Serendipity plays a role in every creative act: a fortunate accident that meets with sagacity (keenness of insight), enables the natural impulse to action, and encourages a playful, non-judgmental attitude toward tinkering and experimentation. An organizational acceptance of serendipity will foster “creative redundancy”—unused human potential for change—a vital reinvesting and self-generating augmentation of the precious human resources within every networked and virtual organization. This meta-systems value can be observed in the real benefits derived from extended employee cross-training and job-rotation, peer sharing of ideas and skills, and the meaningful individual fulfillment and powerful motivation associated with continuing professional development and life-long personal improvement.

**Principle Five: Diverse Stimuli**
- Concepts such as “breaking set”, “cognitive dissonance” and “thinking out of the box” indicate the ostensible value of divergent thinking and novel approaches to creative problem solving. The presence of any stimulus can potentially provide fresh insight into routine tasks as well as new initiatives. Networked and virtual organizations should actively promote exposing a variety of diverse stimuli to its participants and recognize that many creative acts are evoked as a result of novel, challenging, unconventional, or otherwise interesting experiential and environmental stimuli.

**Principle Six: Within-Organization Communication**
- Internal, intra-organizational communication often happens naturally at smaller enterprises but less easily at larger ones since formal communication channels (including routine and standing meetings) are of limited use for supporting creativity. Networked and virtual organizations are particularly well-suited to foster, augment, and facilitate systemic, cross-pollinating, creative, cooperative, and collaborative interpersonal communications. As posited by the effects of serendipity and other unofficial activity—the majority of creative acts are unplanned and unanticipated—informal information exchanges that allow projects that have not been planned to self-organize and incubate. Various channels for improving and supporting communication affordances for members of networked and virtual entities should be thoughtfully implemented.

**CONCLUSION**

In conclusion, superior human factor engineering requires a deep understanding of the theory
and practice of collaborative team leadership and participative stakeholder buy-in, as well as systems theory and the complete, concurrent, iterative IT development lifecycle approaches from needs assessment, task analysis, and conceptual design to production, quality assurance, usability design and testing, and courseware product deliverables and evaluation. Post-modern networked and virtual organizations become successful as their technology systems give their high-tech workers confidence, trust, and a true sense of stakeholder-ship, commitment and collaboration—and value each high-performing team member. Superb management of people and technology within these enterprises needs to support the special cohesive chemistry together with good communication that promotes openness within the organizational idea-generating infrastructure to everyone at every level. Human factor engineering also recognizes the distinction between management and leadership—both of which must become an essential component of team projects and practice. Management skills are essentially functional and structural in nature, and can effectively be performed through the implementation of unambiguous, clearly identified, thoughtfully-designed systems, and organizational approaches to problem requirements. Leadership is another realm altogether: it is essentially a dynamic, protean, flexible, pliant, and highly-adaptable quality associated with effective people skills, awareness of self and others, a coherent, creative vision about both the direction of the product and team as well as the motivated, goal-oriented mission that team members are united in achieving. These are the real “human factors” that underlie all effective leadership...a fundamental understanding that the most precious and irreplaceable resource in any IT business endeavor or enterprise are the human resources at the core of all user-centered, real-world design solutions. In a nutshell, the optimal vision statement for networked and virtual organizations should codify that people come first—both from the perspective within the organization functionally and operationally as well as on “the business end” of the equation as clients and end users of the products and services however defined and delivered.

The emerging applied profession of human factors engineering may make a positive, significant, and possibly decisive change in the way we will design our networked and virtual environments in the 21st century. The human factors and allied usability professions are in the process of improving user experience designs for more accessible, more truly universal design solutions. By simplifying and standardizing IU design affordances, metaphors, and mnemonics, reducing seldom-needed collateral features, incorporating more intuitive, familiar, parsimonious, and easily grasped mental models to reduce working memory loads while improving simple functionality—we can empower all users to “take control” over their technologies...and their networked and virtual, knowledge-centered, and learning organizations!

REFERENCES


Human Factors for Networked and Virtual Organizations


KEY TERMS

Amberfication Paradox: Static information resources, media representations, and human-technology communication systems become non-dynamic as soon as they are produced. The paradox is that the more quickly these fixed entities are produced—the quicker they become conceptually and substantively sealed in amber and disconnected from modification, alteration, correction, and change (i.e., off the information grid) as utilities and resources.

Computer-Human Interface Design: Computer-human interface design consists of the effective functional synthesis and operational integration of three overlapping sub-design domains: information design, interaction design, and media design (See additional definitions of these terms in this article section). Each of these three layers must be clearly defined, represented, and manifested in the user interface design solution. These three sub-design realms/components are synthesized holistically by the conceptual usability glue of the principles of universal design. User interface affordances can be evaluated through usability heuristics and the degree to which these principles have been effectively applied and realized in the user experience—providing a coherent, logically-consistent, and robust (i.e., intuitive, accessible, easy-to-learn, and use) systems design.

Human Factors Engineering: Human factors is an emerging applied design field that entails a synthesis of cross-disciplines from classical ergonomics (i.e., the science of work, system-person interaction, and functional/operational performance needs), and newer research gleaned from the allied fields of cognitive science, human physiological psychology, perception, learning, memory and brain behavior science, interaction/interactivity design, product design, media design, communication design, and information design. Human factors engineering recognizes (1) the intensely cross-disciplinary nature of the human-technology system designer’s role and the transcending cultural, economic, societal, and geopolitical importance of information utilization and knowledge generation, and (2) the field’s creative, protean, and dynamically-hybrid “human-centered, user-centered, and usage-centered” systems approaches that strives to integrate existing multi-disciplinary domains (i.e., change management, computer and information science, informatics and systems theory to management philosophy, law and ethics, human cognition and perception, human resources, usability design, industrial design, and ergonomics).

Information Design: Understanding and clearly mapping the full scope, sequence, concepts, principles, examples, and underlying inheritance structure of the declarative and procedural facets of the content domain (i.e., the infosphere). Information design is essentially concerned with understanding the purpose, the organization, the context, and interrelationships within a knowledge domain: the “envisioning” of information to be communicated.

Interactivity Design: Understanding the affordances of the computer-human interrelationship from the user’s perceptual/cognitive experience; identifying and specifying the dynamic time/space transactions between the user, information, and media elements—creating a graphical user—environmental interface to navigate within that is intuitive, comprehensible, robust, and engaging. The core artifact of interactivity design is the user interface. Principal ingredients of the interface design are signage, cueing, mnemonics, style/layout conventions, and the ambient conceptual metaphor.

Media Design: Media design involves the physical, functional, and operational manifesta-
The media design is the tangible, concrete, tactile-audio-visual-sensori-motor experiential “front end” of the human-technology system. Understanding the multi-sensory nature of the user’s experience; applying human learning, memory, messaging, and perception/cognition to produce effective, aesthetic multiple media that provide cognitive/perceptual/physical affordances to improve human-machine system communications for specific audiences and organizational requirements.

**Usability:** The degree to which human-technological systems, artifacts, and products are appropriately and efficiently designed for the user (i.e., ease of use) is the indication of that product’s usability. Various heuristics and criteria can provide an objective, empirical basis against which to measure and evaluate the level and degree of design efficiency corresponding to the key construct of usability.

**Usage-Centered Design:** Usage-centered design focuses primarily on the functional goal-based behavior of learners and structuring activities, procedures, processes, and corresponding affordances to optimize the effectiveness of the learner to efficiently accomplish those intrinsic goals. In both user-experience and usage-centered approaches, however, the conventionally-deterministic structure of the communication content and the underlying information architecture of the knowledge domain are secondary considerations while the user’s own conceptual model and intrinsic goal-driven behaviors provide the guiding blueprint for the system-product design solution.

**User-Experience Design (UXD):** User-centered (a cognitive/perceptual term) and usage-centered (a behavioral/functional term) are post-modern design descriptors often arbitrarily or ambiguously defined and interchangeably used and mis-used. In the context of 21st century communication product design theory and practice, user-centered design focuses on constructing a user experience and environment with physical and virtual affordances that are manipulable, controllable, customizable, and adaptable from the essential perspective of the conceptual model of the user. This means both (a) the user’s own internal metamodel of their own goal-directed processes, activities and contextual (i.e., socio-psychological and physical) environment, and (b) the designer’s representational model of the user-activity-environmental experience, with the former driving and superseding the latter in the design solution. Thus the conceptual model of the learner becomes the superordinate principle guiding the design process and learning outcomes (i.e., the highest level of the prescriptive taxonomy).

This work was previously published in the Encyclopedia of Networked and Virtual Organizations, edited by G.D. Putnik and M.M. Cunha, pp. 668-677, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 6.9
Trust in Computer-Mediated Communications: Implications for Individuals and Organizations

Susan K. Lippert
Drexel University, USA

ABSTRACT
This chapter explores the concepts of trust as they relate to computer-aided communications. The author defines trust for interpersonal and for inanimate technology, identifies common factors affecting quality of meaning in communication, and identifies ways businesses can foster and enhance trust independent of the medium of communication.

INTRODUCTION
The purpose of this chapter is to explore trust as it relates to computer-aided communications, commonly electronic-mail. There are four objectives for this chapter: (a) to comprehensively define trust as it appears in the scholarly literature, both for human interpersonal trust and for inanimate technology trust; (b) to explore the essence of communications, identifying the common factors that affect the quality of meaning in communicative interactions; (c) to compare and contrast trust in different forms of communication media; and (d) to offer some thoughts on what can be done in the business world to foster and enhance trust, independent of the communication medium chosen.

TRUST
Trust is a contextual phenomenon commonly applied to casual conversation without conscious knowledge of what the construct means or how it manifests in daily interactions. This suggests that trust is a latent variable in the communications process since many people are not conscious that what they say and how they say it can affect their
Trust in Computer-Mediated Communications

trust relationships with others. Scholars often debate how to operationalize trust and, consequently, how to measure whether an individual displays and demonstrates trust. While the definitions of trust vary, there is agreement that its meaning is situationally or contextually based (McKnight & Chervany, 1996; McKnight, Cummings & Chervany, 1998), which further suggests that how trust is operationalized is partly a function of the object upon which trust is being placed.

Rotter (1971, p. 443) suggested that “the entire fabric of our day-to-day living, of our social order, rests on trust—from buying gasoline, paying taxes, going to the dentist, flying to a convention—almost all of our decisions involve trusting someone [or something] else.” Other scholars have noted that trust is central to all social transactions (Dasgupta, 1988), reduces transaction effort (Bromiley & Cummings, 1995), and is classified as an important component of social systems (Arrow, 1974). Trust has been cited as a vital form of social capital within social systems (Coleman, 1990; Fukuyama, 1995), since “without trust … everyday social life … is simply not possible” (Good, 1988, p. 32).

Much of the trust literature, particularly in the area of organizational theory and management, has focused on interpersonal trust where the object of trust is another individual. This form of trust, interpersonal trust, is most commonly defined using the research of Mayer, Davis, and Schoorman (1995, p. 712), who suggest that interpersonal trust is:

… the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.

Other definitions frame trust as “a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another” (Rousseau, Sitkin, Burt, & Camerer, 1998, p. 395). The trust that one individual places in another may fluctuate during the course of a relationship based on a variety of external stimuli. That is, trust is not a naive faith that a party takes for granted, based on an interaction that occurred in the distant past (McEvily, Perrone, & Zaheer, 2003). Instead, individuals examine new information about those with whom they interact and decide if they should increase or decrease their trust in that individual (McEvily et al., 2003). Using Mayer et al.’s (1995) interpersonal trust definition, we can state that the evaluation of another’s trustworthiness is a function of three antecedents: ability, benevolence, and integrity. The key difference between trustworthiness and trust is that trustworthiness is a perceived characteristic of the trustee, while trust is a psychological state of the trustor (Sparito & Lippert, 2006). Trust becomes relevant when individuals develop dependencies on, and vulnerabilities to, the actions and decisions of others (McEvily et al., 2003).

Levin, Whitener, and Cross (2006) found differences in an individual’s willingness to trust another when they classified relationships into new, intermediate, and older relationships. In their research, they also found that in newer relationships, the basis for trust was gender parity, perhaps as a function of communication or personal style, that behavioral expectations that result from moderate social interaction affected intermediate relationships, and that a personal knowledge of shared perspectives (values, beliefs, perceptions, and environments) was linked to older relationships. This suggests that how individuals behave and reinforce trust, by communicating, may serve as a moderator for trusting relationships. Zahra (2003, 2005) suggested that familial and kinship ties encourage trust as a function of the depth of the relationship. In a recent study, Zahra, Yavuv, and Ucbasaran (2006) suggested that managers can build trust relationships with individuals and groups through solicitation of ideas, problems, and questions. This communication strategy has
the capacity to enhance trust relationships in an organizational setting.

In an empirical investigation of 18 software vendor companies based in India, Oza, Hall, Rainer, and Grey (2006) found that communication was a key component necessary to maintain trust in established outsourcing relationships. In fact, Oza et al. (2006, p. 352) reported that one of their interviewees stated that “there should be enough communication between [the] vendor and client to prosper trust in the relationship.” Henttonen and Blomqvist (2005) proposed that communication behavior that includes timely responses, open communication, and providing useful feedback were considered to build trust in virtual teams. Lewicki and Bunker (1996) along with Mayer et al. (1995) assert that trust-building is an experiential process that occurs over time as individuals engage in continual and repeated conversations. Social communication such as exchanging greetings, interests, and other personal information is important to the evolution of trust in both virtual and traditional teams (Järvenpaa & Leidner, 1998). Henttonen and Blomqvist contend that in the early stages of a relationship, both the content (social vs. work-related) and context (face-to-face vs. intranet, or e-mails) should be mixed rather than being treated separately because the combination of communication mechanisms appears to have a strong tie and connection to emotional commitment and trust.

Pepper and Larson (2006) discuss differences in the use of information communication technologies (ICTs) and face-to-face communication in geographically dispersed organizations that were recently acquired. They identified four concerns regarding the use of ICTs in organizational settings as shown in Table 1.

<table>
<thead>
<tr>
<th>Concern</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Improper ICT use exacerbates commitment problems</td>
</tr>
<tr>
<td>2</td>
<td>Cultural differences influence technology use</td>
</tr>
<tr>
<td>3</td>
<td>Over-reliance on ICTs can lead to a lack of trust</td>
</tr>
<tr>
<td>4</td>
<td>The timing of ICT choices influences employee perceptions</td>
</tr>
</tbody>
</table>

Each concern demonstrates that the use of technology to facilitate communications may indeed result in unexpected outcomes such as a paucity of interpersonal trust. What one can learn from these concerns is that managed use of ICTs is important if trust between employees is to be maintained and nurtured. Trust may form in relationships in which interpersonal contact ranges from extensive to nonexistent (Wilson, Straus, & McEvily, 2006).

The above studies suggest that what is said, how it is said, and the media used to facilitate communication between two individuals in an organizational setting will serve as moderators to the trust relationship. Therefore, conscious and thoughtful reflection on these elements, prior to and during the communication process, should offer organizational members the opportunity to facilitate trust in their interactions.

While trust between individuals has received the most attention in the scholarly literature, other forms of trust exist that also require investigation. Research has shown that the object of trust determines the type of trust relationship under
Trust in Computer-Mediated Communications

consideration (Giffin, 1967). Trust may develop between individuals (Johnson-George & Swap, 1982) in both professional and personal settings. Additionally, trust can also develop between individuals and organizations (Zaheer, McEvily, & Perrone, 1998); between organizations (Gulati, 1995); between individuals and social institutions (Barber, 1983); between individuals and technology (Lippert, 2001, 2007, Lippert & Davis, 2006; Lippert & Forman, 2006; Lippert & Swiercz, 2005); as a general characteristic of different societies (Fukuyama, 1995); and as a personal trait (Rotter, 1971). The principle differentiator between the forms of trust is the object of trust, which Giffin (1967), asserts can be a person, place, event, or object.

Trust has been investigated in many social sciences including management, economics, information systems, psychology, sociology, political science, anthropology, organizational theory, and communication. Although not all inclusive, Table 2 offers a view of disciplines in which the trust phenomenon is studied.

Technology Trust: An Alternative Object of Trust

Technology trust is defined as:

\[
\text{the extent to which an individual is willing to be vulnerable to the information technology (IT) based on expectations of technology predictability, technology reliability and technology utility and influenced by the individual’s predilection to trust technology. (Lippert, 2001, p. 9)}
\]

Trust assessments are often based upon a single interaction with a technology and then reinforced or diminished each time an information system is used (Denning, 1993). If the technology is operational when needed, a positive assessment of system performance is recorded. Users may

<table>
<thead>
<tr>
<th>Disciplines</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management</td>
<td>Ferrin, Dirks &amp; Shah, 2006; Roy &amp; Dugal, 1998; Kramer &amp; Tyler, 1996</td>
</tr>
<tr>
<td>Economics</td>
<td>Bertrand, Duflo &amp; Mullainathan, 2004; Dasgupta, 1988</td>
</tr>
<tr>
<td>Information Systems</td>
<td>Bekkering &amp; Shim, 2006; Lippert, 2001, 2007; Lippert &amp; Davis, 2006; Lippert &amp; Forman, 2006; Lippert &amp; Swiercz, 2005; Muir, 1994; Gefen, Karahanna &amp; Straub, 2003; Oza et al., 2006</td>
</tr>
<tr>
<td>Psychology</td>
<td>Overwalle &amp; Heylighen, 2006; Rotter, 1967, 1971</td>
</tr>
<tr>
<td>Sociology</td>
<td>Sitkin &amp; George, 2005; Welch, Rivera, Conway, Yonkoski, Lupton &amp; Giancola, 2005; Zucker, 1986</td>
</tr>
<tr>
<td>Political Science</td>
<td>Barber, 1983; Leach &amp; Sabatier, 2005; Letki &amp; Evans, 2005</td>
</tr>
<tr>
<td>Anthropology</td>
<td>Blum, 1995; Carrithers et al., 2005</td>
</tr>
<tr>
<td>Organizational Theory</td>
<td>McKnight et al., 1998; Lewicki &amp; Bunker, 1996; McAllister, 1995; Mayer et al., 1995; Wilson et al., 2006</td>
</tr>
<tr>
<td>Communication</td>
<td>Coppola, Hiltz &amp; Rotter, 2004; Hubbell &amp; Chory-Assad, 2005; Walther &amp; Bunz, 2005</td>
</tr>
</tbody>
</table>
consider frequent or inconvenient downtimes as negative experiences with the system. Past experiences with the technology, both positive and negative, influence an individual’s assessment of that system as a whole.

This suggests that the use of technology to facilitate communication may offer some additional insights into what causes individuals to develop or lose trust in others based on exchanges undertaken via ICTs. An investigation into trust in technology has the potential to offer further insight into whether a user’s trust of a specific technology is likely to impact use of that system. This suggests that dependence upon technology for the completion of daily tasks makes individuals vulnerable to the technology. Employees depend upon technology to process financial transactions, maintain corporate Web sites, schedule meetings with clients, and communicate with geographically dispersed coworkers. However, computers are fallible, experience downtimes, and sometimes fail to function consistently from day to day. Depending upon the task, the lack of functional technology can, at a minimum, temporarily disrupt employee performance, or, at worst, put a halt to any productivity for an indeterminate length of time. These potential problems may have an impact on employees’ willingness to use the technology to communicate information resulting in possible trust issues.

THEORETICAL BACKGROUND

Theory Overview

Proposed by Daft and Lengel (1984, 1986), media richness theory (MRT), also known as information richness theory, establishes a scale to rank communication media based on the quality, breath, and depth of information in order to explain managerial media selection behaviors. MRT research is a foundation for later media selection theories and a direct base for modern communication theory. MRT explains how organizations contend with uncertainty and equivocality, or ambiguity of interpretation within organizational communication.

Classification Criteria of Media Richness Theory

Media richness theory asserts that individuals use media selection to reduce ambiguity in communication. Richness, in this context, is the ability of the communication channel to not only transfer data but also, more importantly, to impart meaning. Richness is achieved by carrying equivocal information, which modifies the communication participants’ understanding. Understanding occurs when different conceptual frames of reference converge or ambiguous issues are resolved. If a particular medium provides new understanding or carries equivocal information effectively, the medium is deemed rich; otherwise, it is considered a lean medium.

Organizational tasks differ in terms of ambiguity (Daft & Macintosh, 1981) and communication reduces task ambiguity, thereby providing alternative solutions to a given problem (Guinan & Faraj, 1998). Likewise, communication media vary in their ability to reduce ambiguity in communication. According to MRT, ambiguity is reduced through a blending of four criteria based on the communication medium’s ability to: (a) facilitate feedback; (b) convey meaning through multiple cues; (c) use a variety of language; and (d) present personalized messages. Medium richness results from this blending of criteria used to rank media along a richness continuum. “Rich” media reduce high levels of ambiguity whereas “lean” media are sufficient for tasks of low ambiguity.

Communication Media and Information Richness

The classification criteria in Table 3 operationalize qualities used to assess media richness.
Communication media with more classification features rank higher on the richness scale than media with fewer richness features. The level of understanding between the communication participants determines the richness of the communication. Daft and Lengel (1984, 1986) used these criteria to rank five communication media on a richness continuum (Table 4). Media, in richness identity, rank from high to low: face-to-face, telephone, personal written text (letters, memos), formal written text (documents, bulletins), and formal numeric text (computer output). Although not included in early studies, electronic mail was later ranked between telephonic communications and written personal documents in the context of richness (Daft, Lengel, & Trevino, 1987).

Face-to-face communication ranked highest in information richness for two reasons. First, face-to-face communication provides for multiple cues (verbal and nonverbal) and second, offers instant feedback between communication participants. In face-to-face communication, a high degree of message personalization is supported.

Table 3. Media richness classification criteria compiled from Daft and Lengel (1984, 1986)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback</td>
<td>Instant vs. delayed feedback—offers an opportunity to ask questions and make corrections</td>
</tr>
<tr>
<td>Multiple cues</td>
<td>The capacity to convey meaning through multiple cues including visual cues, audio cues, body language, tone of voice, facial expression, words, numbers, and graphic symbols</td>
</tr>
<tr>
<td>Language variety</td>
<td>The capability to customize the message by using different words to increase understanding</td>
</tr>
<tr>
<td>Personal focus</td>
<td>The extent to which a person can convey personal or impersonal feelings in the communication</td>
</tr>
</tbody>
</table>

Table 4. Communication media and information richness compiled from Daft and Lengel (1984)

<table>
<thead>
<tr>
<th>Communication Media</th>
<th>Information Richness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face-to-face</td>
<td>Highest</td>
</tr>
<tr>
<td>Telephone</td>
<td>High</td>
</tr>
<tr>
<td>Electronic mail *</td>
<td>Moderate</td>
</tr>
<tr>
<td>Written, Personal (letters, memos)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Written, Formal (bulletins, documents)</td>
<td>Low</td>
</tr>
<tr>
<td>Numeric, Formal (computer output)</td>
<td>Lowest</td>
</tr>
</tbody>
</table>

* Added by Daft et al. (1987)
The telephone offers fast feedback, and audio communication is available via the telephone. Visual and nonverbal communications including hand, facial, or body gestures are non-existent with this medium. However, nonverbal cues derived from voice inflection, tone, and speaking style still are active in telephonic communications. Message personalization is available, although not to the degree of face-to-face communication. Communication is dependent upon language content and audio cues for understanding rather than written word or behavioral cues.

Personal written letters or memoranda offer limited visual cues. Feedback is slow and dependent upon receiver response time. Message personalization is available although less than telephone and face-to-face communication. Meaning must arise from the written word without the benefit of the ancillary nonverbal cues available in richer media. Audio cues are absent and richness is conveyed only through words on the page and the structure of the written word.

Formal written documents such as bulletins are less rich, since feedback is very slow, audio cues and nonverbal communications are absent, and visual cues are limited. Richness is conveyed through written words from which meaning must arise. The nature of this type of communication is often viewed as impersonal.

Formal numeric documents are considered lean. Computer generated reports provide text and numbers which offer limited visual cues. Neither audio nor nonverbal cues are available to enhance understanding. Feedback is very slow and communication is seen as impersonal.

COMMUNICATIONS

A Definition of Communication and Medium

McCroskey and Richmond (1999, p. 7) offer a definition of communication as “the process by which one person stimulates meaning in the mind(s) of another person (or persons) through verbal and nonverbal messages.” Verbal messages consist of language where language is “a set of symbols or codes used to represent certain ideas or meanings” (McCroskey & Richmond, p. 7). Nonverbal messages are any messages other than verbal such as tone of voice, vocal pitch, body position, eye movements, hand gestures, and facial expressions. The authors also emphasize the importance of meaning rather than the physical exchange of messages (McCroskey & Richmond, 1999, p. 7). For purposes of this discussion, their definition is appended to include the notion of an appropriately rich communication channel. As such, the definition of communication becomes:

... the process by which one person stimulates meaning in the mind(s) of another person(s) through verbal and nonverbal messages via an appropriate communication channel.

This definition recognizes the role that the communication channel plays in human exchanges. The channel can be technologically based, such as with electronic-mail, or the channel can be face-to-face contact. This definition encompasses both the presence of technology and the human element. Noise is a mediating factor in the communication channel since it has the capacity to interrupt the message resulting in distorted meaning. For example, in electronic messaging, slow speed might impede the message and meaning that is intended to be communicated.

A Model of Communication

There are various models and diverse definitions of communication. Borden (1971) introduced eight different schemas including a behavioral communication and telecommunications schema. Later models of communication incorporated a “medium” or communication channel. George and Jones’ (1996) model of communication is
used to explain the process of communication, since it allows for the potential influence of the communication medium on the process.

The George and Jones (1996) model includes a communication medium in both the initial transmission and the feedback loop. The message represents the information the sender wants to share with another individual, group, or organization. The message consists of the sender’s thoughts, either conscious and/or unconscious, and can be any form of data or information. Through the encoding process, the sender expresses thoughts in symbols or language, through either written or oral form to the receiver.

**The Media of Communication**

The medium is the mechanism through which the communication is transmitted. The presence of a computer-mediated medium such as electronic-mail creates an artificial communication barrier between the sender and the receiver. This potential message distortion may be either a technological, behavioral, or interpretational error. If the data being exchanged between the sender and receiver is unclear for any reason in the transmission process, a technological error occurs. A behavioral error exists when the actions of the sender and/or receiver are distorted. The interpretational error occurs if the receiver misinterprets the intended message. This medium, regardless of the type, sets up the possibility of additional noise during transmission.

The degree of understanding shared between what the sender intended and what the receiver decoded measures communication effectiveness. Frequently, miscommunication results from noise distortions between the sender’s intent and the receiver’s understanding of the message. Noise may be positive or negative and may include a history of previous communications with the sender, the receiver’s history in decoding communications of this type, the receiver’s current mental state, the method of communication used to convey the message, or the receiver’s health. Both the sender and the receiver construct or deconstruct the message as influenced by noise.

A challenge occurs in ensuring that the message sent and message decoded are congruent. The congruity problems might include: (a) coding and decoding problems (Moorhead & Griffin, 1995); (b) lack of common experience—a lack of common experience occurs when there is no shared language between the sender and the receiver. A shared language experience arises from the use of mutually understood symbols; (c) semantics—semantic problems occur when people attribute different meanings to the same words or language forms; (d) jargon—jargon is specialized or technical language that is specific to a field or profession. Jargon is usually a hybrid form of the standard language where words hold special and atypical meaning to the members of the group; and (5) medium problems—medium problems arise from selecting an inappropriate communication medium for the message being communicated.

Another issue in the communication process is the assumption that the receiver accurately decodes the message (Moorhead & Griffin, 1995) and nothing interferes with the message translation. This problem set might include: (a) selective attention—the problem of selective attention occurs when the receiver focuses on only selected parts of a message; (b) value judgments—value judgments occur when a valiance is place on the message received. If the message corresponds to the receiver’s personal beliefs, the receiver may accept the message without reservation; otherwise, the receiver may disregard the message in its entirely; (c) lack of source credibility—if the sender is viewed as incompetent, unknowledgeable or lacking in credibility, the receiver may partially or completely disregard the message. Likewise, if the source is an expert in the field, the receiver may accept the message without
question. Vital information may be discounted if the receiver questions the sender’s credibility; and (d) overload – communication overload occurs when an individual receives a greater amount of information than they can reasonably process. Organizations are overwhelmed with information being disseminated, such as computer printouts, electronic mail, and voice mail messages.

The medium of communication becomes a complex variable when assessing the quality of a message. In the modern information age, the use of technological systems to transmit and decode messages places an additional burden on the individuals involved in the process. These burdens include the recording, organizing, and interpreting the message.

**The Importance of Effective Communication to Facilitate Trust**

In the communication process, trust becomes an ever increasingly more important phenomenon. Problems such as selective attention, value judgments, lack of credibility, and information overload become manageable elements in the development of trust in an online environment. Trust is enhanced when the receiver is encouraged to attenuate the entire message. Recognition of value judgments as untested evaluations contributes to communication errors and can negatively affect trust. Credibility is closely aligned with trust in interpersonal exchanges in which the receiver views the sender as lacking in source knowledge or competency. Trust becomes an analog to credibility. Information overload becomes extraneous noise that affects the communication process by having the receiver attend to too much information. The intended outcome of the communication process is the development of shared meaning which enhances the predictability and reliability of the information exchange.

**Enhancing Trust in Communications**

Trust is a concept that permits the formation of perceptions based on the degree of predictability, reliability, and utility an individual exhibits toward a communication medium and the degree of faith in the integrity, honesty and benevolence of others. Trust is an underlying precept for all social interactions within both personal and professional relationships. Perceptions of trust and assessments of trustworthiness are oftentimes unconscious and frequently misplaced when individuals use erroneous data to evaluate an exchange with another. The potential noise that can occur, as a function of using a computer-mediated communication mechanism, can generate additional distortions to an otherwise clear exchange of ideas and intentions. In the process of seeking clarity, individuals must guard against making judgmental trust errors, since adjustments in trust perceptions, as a function of these errors, may result in degradations of trust rather than a refinement of trust accuracy. Interactions that are trust dependent serve as a continuous learning device to upgrade or downgrade the trustworthiness dimensions.

One way to minimize these errors is to select a communication medium based on its richness that is matched to the content of the intended exchange. More ambiguous data should be conveyed through a richer communication medium to help support the other’s interpretations of the original intentions. Conscious selection of media to convey rather than using an ICT because it is convenient is a simple process to assist with noise reduction.

Communications is one of the foundations of social interaction. Humans have developed sophisticated systems to convey thought, feeling, and meaning. In the technology age, the evolution of technological systems permits the development and use of mechanical and electronic devices to aid in the speed, clarity, and meaning that is con-
veyed between two individuals. Communication technologies are subject to their own trust evaluation through the assessments of predictability, reliability, and utility offered by the intended user. Hence, trust in technological communication systems is both important and profound since it may offer further explanations for the presence of noise, the potentials for miscommunication, and the unintentional affects on trust. Each use of a system generates an experience that serves as the basis for an individual’s trust judgments about that technology.

The challenge in the modern information age is to be able to trust the conveyed meaning shared with another and to accurately recognize meaning from others, in order to improve the efficiency and effectiveness of message dissemination. Technology provides an added challenge to trust by introducing an alternate dimension, which links our perceptions and judgments about the predictability, reliability, and utility to a communications system. An objective to improve both interpersonal and technology trust through conscious recognition of the factors that affect the evaluation of trust are oftentimes difficult to operationalize. However, provided below are a few recommendations for managers and their employees to facilitate accurate communication that may lead to enhanced trust:

1. Include trust as a conscious consideration in all forms of communication rather than something that is considered when a problem in communication occurs;
2. Recognize that trust perceptions of people and technology are oftentimes distorted and in error;
3. Consistently work to say what we mean and then mean what we say;
4. Conduct continual assessments of the qualities of the communications and change the medium, the message, or the delivery to optimize transfer of meaning;
5. Clarify as many variables as possible when a technological system is used to convey a message;
6. Select media based on richness to reduce ambiguity in communications; and,
7. Take full advantage of modern information technology to aid in the communication process by using technology where feasible and desired.

FUTURE TRENDS

Sometimes the popular press suggests that we have reached the limits of our capability to effectively communicate, whether by direct human interaction or through new and more complex technology. The future, however, holds promise for the enhancement of communication styles, structures, patterns, and practices, through greater understanding of the many factors that affect the communication process. We can expect to see improvement in the quality of workplace communication, through the use of more advanced and sophisticated technologies to help the message get through accurately, effectively, and efficiently. At the same time, the workplace recognizes that because communication is the mechanism in which business is conducted, better understanding, improved skills, and more sophisticated technology will all contribute to increase capacity. The breath and depth of communications will increase and the technological system used to assist in the operationalization of information sharing will become more accurate, user friendly, predictable and reliable, and more useful, leading to an ever-increasing trust which aids human interaction. As we become more conscious of the affect of trust on communications, we need to continue the development of our knowledge, skills, and attitudes toward improved practices both within and outside the work environment.
REFERENCES


Giffin, K. (1967). The contribution of studies of source credibility to a theory of interpersonal
trust in the communication process. *Psychological Bulletin*, 68(2), 104-120.


Lippert, S. K. & Swiercz, P. M. (2005). Human resource information systems (HRIS) and tech-


*This work was previously published in Computer-Mediated Relationships and Trust: Managerial and Organizational Effects, edited by L. L. Brennan and V. E. Johnson, pp. 1-13, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 6.10
Understanding an ERP System Implementation in a Higher Education Institution: A Grounded Theory Approach

Jose Esteves
Universitat Politecnica de Catalunya, Barcelona, Spain

Joan Pastor
Universitat Internacional de Catalunya, Barcelona, Spain

ABSTRACT

This chapter addresses the problem of enterprise resource planning (ERP) implementation in Higher Education Institutions (HEI). It attempts to contribute to the understanding of ERP implementations in this kind of organizations by identifying and analyzing the major factors that affect this type of projects. Special attention has been paid to contextual influence and to organizational factors. The conclusions of this work suggest that most problems in ERP implementation projects are not technological but may be attributed to organizational factors. The chapter describes an in-depth case study carried out at a HEI that implemented an ERP system in 2001. The case was studied as part of a grounded theory based research project whose aim was to develop a model for the implementation of ERP systems.

INTRODUCTION

In recent years, a growing number of HEIs worldwide are exploring the use of ERP systems as a means of supporting their organizational processes, while linking areas like financial, real estate, and staff management, management of students, and support of teaching and learning. This adoption of ERP systems by HEIs is bringing problems in ERP implementation projects that are specific to these types of organizations. Because HEIs comparatively have modest IT/IS resources and budgets, they expect to benefit largely from
implementing off-the-shelf application products like ERP systems. They do not usually have the depth of experience or a constant availability of adequate expertise to be able to handle the in-house development of an enterprise-wide system.

This chapter reports the results of an in-depth case study carried out in the economics management area of a Spanish HEI that implemented an ERP system in 2001 by following a “big-bang” implementation approach. The big-bang approach is characterized by the “go live” of all implemented ERP modules at the same time.

The interpretive perspective adopted in our research reflects our aim for understanding the phenomenon of ERP implementation in a HEI within the full-fledged organizational context where it occurs. In order to keep the confidentiality of the research site and the people involved in this case study, we deliberately omitted the name of the institution and ERP product. The chapter is structured as follows. First we present the literature review on ERP. Then we describe the case study background. Next, we present the research methodology. Subsequently, we describe the findings. Finally, the conclusions and the implications for further research are outlined.

PRIOR RESEARCH ON ERP

Esteves and Pastor (2001) reviewed 189 articles on ERP topics that were published in conferences and top information systems journals. They noted that implementation approaches have been studied and that new ones have been proposed. However, they detected that “implementation” does not mean the same to all authors. Specifically, in what concerns the process, different authors consider different models with different phases and stages. The survey findings suggested that research on the critical factors relevant to ERP implementation success was covered with some studies that did not provide precise definitions of those factors found. Therefore, more effort should be put into defining ERP implementation critical success factors and their subsequent validation.

Aladwani (2001) stated the following:

Past ERP implementation research may be described as factor research, which involves identifying the factors or variables that are critical for implementing ERP successfully. Although factor research is valuable for advancing our understanding of ERP implementation success, it adopts a static view, which limits its adequacy in explaining the dynamics of the implementation process. (p. 267)

Aladwani (2001) and Robey, Ross, and Boudreau (2002) suggested a process research approach or a combination of factor and process approaches, in order to improve research in ERP topics. Using a process approach, ERP implementation may be conceived as sequences of discrete events that lead to outcomes of particular interest. Some studies focused on ERP impacts at organizational, technological, and business levels, business process reengineering, and organizational change management issues. However, the few studies developed so far are not enough to create a body of knowledge in the area. Case studies, not necessarily in-depth, constituted the largest category of articles. From the perspective of research methodology, Esteves and Pastor (2000) detected that in some of the referenced studies, there was a lack of explanation of the research methodology applied, lack of data to support the presented results, and lack of assumptions or hypotheses (in theoretical terms) for further studies. Furthermore, although some studies attempted to identify critical success factors for ERP implementations in general (e.g., Esteves & Pastor, 2000), and more specifically to HEI (Nielsen, 2002), there is a lack of research on the management of these critical success factors.

In sum, as Robey, Ross and Boudreau (2002) mentioned, the research on ERP has been mostly descriptive. They pointed out that “little atten-
tion has been paid to developing a compelling theoretical explanation of ERP implementation, which is needed to explain contradictory findings and to allow generalization of findings to related phenomena” (p. 21). Due to lack of theoretical foundations on the ERP domain, it seems appropriate to investigate ERP implementations by using grounded theory applied to data drawn from in-depth case studies.

CASE STUDY

The Spanish HEI of our case study defines its mission in terms of quality teaching, research, and technology transfer. It was created in the 1970s, and it is composed of several academic schools and research institutes with more than 30,000 students and more than 2,000 teachers.

The HEI economics management area is composed of two general departments: the economics department and the technology transfer unit. There are also economics management personnel in each structural unit. Until 1999, both the economics department and technology transfer unit had their own legacy systems that were totally independent. The units (schools and institutes) were responsible for sending the accounting documents to the economics department or technology transfer unit, and these departments processed them in the respective legacy systems. Table 1 describes the different actions taken to improve information technology infrastructure in this HEI.

This HEI followed a typical big-bang ERP implementation approach with a well-known European ERP package. The number of expected end users was around 220. The HEI implementation project followed the typical implementation phases: planning, design, realization, go live, and support. At the beginning of the project, it was estimated that some be-spoke development would be made to fulfill some special requirements.

This ERP project had a delay of 1 year in relation to the initial project plan. Based on a process research approach, we describe the main events and activities that may have caused this delay.

---

Table 1. Information technology evolution in HEI

<table>
<thead>
<tr>
<th>Period</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Until 1998</td>
<td>• Diverse legacy systems for different purposes in Economics Management</td>
</tr>
<tr>
<td></td>
<td>• No connection between the legacy system</td>
</tr>
<tr>
<td></td>
<td>• Different economics management procedures within the Economics Management area</td>
</tr>
<tr>
<td>Beginning 1998</td>
<td>• Analysis of legacy systems upgrade or adoption of an ERP system due to Y2K problems, Euro conversion, and very high maintenance costs.</td>
</tr>
<tr>
<td></td>
<td>• Selection of an ERP system</td>
</tr>
<tr>
<td>End of 1998</td>
<td>• Decision to adopt the ERP system.</td>
</tr>
<tr>
<td></td>
<td>• Selection of the ERP implementation consulting company.</td>
</tr>
<tr>
<td>1999-2001</td>
<td>• Beginning of ERP implementation project.</td>
</tr>
<tr>
<td></td>
<td>• Adaptation of the information technology infrastructure.</td>
</tr>
</tbody>
</table>
RESEARCH METHODOLOGY

As we believe that the understanding of ERP implementations cannot be achieved without considering the organizational context in which they occur, the main research method chosen was in-depth case study (Yin, 1994). Furthermore, we opted for an interpretive research approach. Interpretive research does not predefine dependent or independent variables, and it attempts to explain phenomena through the meanings that people assign to them (Orlikowski & Baroudi, 1991).

We started the case study by defining a data collection plan. The main technique chosen for data collection was semistructured interviews. The HEI president (vice president at the time of the ERP implementation) and the ERP project manager were initially contacted in order to permit the study to be carried out and to collaborate in this research project. After permission was obtained, the project manager granted access to most of the documents produced during the project. Later, we asked to visit and interview them, as well as other people who played relevant roles (external consultants, project team members, key users, end users) during the ERP implementation project.

Our first task was to analyze the documentation created during the ERP implementation project. This documentation helped us to understand the project background and to prepare the questions for the interviews. Later, data from interviews were triangulated with the documentation so far accumulated. In order to build theory from this case study, we decided to adopt the grounded theory method.

Grounded theory is a general method developed by Glaser and Strauss (1967) for building theories that are grounded in data systematically gathered and analyzed. Strauss and Corbin (1990) explained the use of grounded theory:

A theory is inductively derived from the study of the phenomenon it represents. In Grounded Theory method one does not begin with a theory, and then prove it. Rather, one begins with an area of study and what is relevant to that area is allowed to emerge. (p. 23)

In our study, the coding process of all interviews and documentation allowed for major themes and categories to emerge. Then, we used the paradigm model proposed by Strauss and Corbin (1990) to relate these categories. Briefly, the paradigm model encompasses the following elements: causal conditions, phenomenon, context, intervening conditions, strategies and actions, and consequences.

A Grounded Model from ERP Implementation

In this section, we describe the specific paradigm model (Figure 1) that we developed by using grounded theory from the data collected from our case study.

Phenomenon

Strauss and Corbin (1990) stated that phenomena are “the central ideas in data represented as concepts” (p. 101). According to their account, the purpose behind naming phenomena is to enable researchers to group similar events, happenings, and objects under a common heading or classification. The phenomenon in the paradigm model is represented by the central category (sometimes called the core category) that represents the main theme of the research. The phenomenon addressed in this study is the implementation of an ERP system in a HEI.

Causal Conditions

Based on the project documents and the interviews, we identified the following reasons to adopt the ERP system: legacy system obsolescence, complex and expensive maintenance process, Y2K problem and Euro conversion, and existence
Understanding an ERP System Implementation in a Higher Education Institution

Environmental Context

By the time ERP implementation at our HEI took place (beginning in 1999), none of the Spanish universities were implementing ERP systems. However, most Spanish universities were thinking about improving their information systems, possibly through abandoning their legacy systems and migrating to ERP systems. Again, problems related to the Y2K problem and the new Euro currency were viewed as strong justifications for carrying out deep changes in information systems. At that time, ERP vendors started approaching the HEI market, and they began to offer ERP solutions adaptable to the HEI market.

The Spanish government also started changing legal requirements and obligations in terms of financial reporting, especially with regard to budget control.

Organizational Context

The organizational culture of our studied HEI is characterized as a bureaucratic culture. According to Wallach (1983), bureaucratic cultures have
clear lines of responsibility and authority and work is highly organized, compartmentalized, and systematic. The information and authority flows are hierarchical and based on control and power. Overall, bureaucratic companies tend to be mature, stable, and relatively cautious.

Our HEI possesses a culture in which most of the times, delays in decision making and expected results are explained in terms of “due to the process,” which refers to the high procedure orientation within this HEI. In this context, it was expected from top management that the ERP implementation would change the way things worked in the organization. However, there was the general perception that organizational culture would not allow for maximum benefits from the ERP system to be achieved, mainly due to the lack of power of some of the involved units.

Intervening Conditions

Because an adequate project manager role is one of the ERP implementation critical success factors, we describe this aspect in the next section - actions and interaction strategies. The HEI started on the selection of the consulting company after the selection of the ERP system. The main arguments for selecting the consulting company were ERP vendor recommendation and experience with ERP implementation in the public sector. At the time of the implementation of the project, this consulting company lacked senior personnel in the ERP implementation area, a fact that was then unknown to the HEI.

Action/Interaction Strategies

In the action and interaction strategies, we used as a priori categories the critical success factors from the unified model first proposed by Esteves and Pastor (2000) and shown in Table 2. Although some studies such as Nielsen (2002) tried to identify critical success factors specific for HEI projects, we opted for our generic model for ERP implementation projects.

Due to space limitations, in this chapter, we focus on the strategic and organizational perspectives within Table 2. The strategic perspective is related to the core competencies accomplishing the organization’s mission and long-term goals. The organizational perspective is related to concerns like organizational structure and culture and business processes. Next, we describe the findings according to each critical success factor within these perspectives.

Sustained Management Support

According to the interviewed stakeholders, in the case of the HEI, the continuous support of the top management was exercised specifically by the then vice president of the university. It was he who participated more actively in the ERP implementation process, mainly after the initial deviation regarding the first project plan. The vice president participated in almost all the meetings of the steering committee. He motivated the project team and contributed with his knowledge about the operation of the university, both at the general level and in terms of its financial management.

The vice president role was crucial as a mediator during some tensions arising from resistance to change. A member of the steering committee manifested that “a project of this type was not high priority for the university.” Initially, top management presented the ERP project like a project bounded for the economics management area; they did not look at it as a strategic project for the whole university.

Effective Organizational Change Management

According to the implementation consulting company, the change management process was in the charge of the HEI, because the HEI decided
Understanding an ERP System Implementation in a Higher Education Institution

Table 2. Critical success factors unified model (Source: Esteves & Pastor, 2000.)

<table>
<thead>
<tr>
<th>Strategic</th>
<th>Tactical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sustained management support</td>
<td>Dedicated staff and consultants</td>
</tr>
<tr>
<td>Effective organizational change management</td>
<td>Strong communication in-wards and outwards</td>
</tr>
<tr>
<td>Good project scope management</td>
<td>Formalized project plan/schedule</td>
</tr>
<tr>
<td>Adequate project team composition</td>
<td>Adequate training program</td>
</tr>
<tr>
<td>Comprehensive business process reengineering</td>
<td>Preventive troubleshooting</td>
</tr>
<tr>
<td>Adequate project sponsor role</td>
<td>Appropriate usage of consultants</td>
</tr>
<tr>
<td>Expert involvement and participation</td>
<td>Empowered decision makers</td>
</tr>
<tr>
<td>Trust between partners</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technological</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequate ERP implementation strategy</td>
<td>Adequate infrastructure and interfaces</td>
</tr>
<tr>
<td>Avoid customization</td>
<td>Adequate legacy systems knowledge</td>
</tr>
<tr>
<td>Adequate ERP version</td>
<td></td>
</tr>
</tbody>
</table>

On the most favorable side, the economics department presented optimistic expectations of the project, not because of the ERP system but because of the obsolescence of its own legacy system. According to some interviewees, they also had the opportunity of taking advantage of the ERP implementation project to carry out an internal reorganization of the unit. On the other hand, some of those unit managers who showed a rejection to the system from the beginning did it mainly for three reasons: they knew of some bad ERP experiences in nonacademic organizations; they were already sufficiently satisfied with the economics management functionality locally offered by their respective dedicated and isolated legacy systems; and they did not see justification for changing their local economics management processes. These three reasons were argued explicitly in the case of the technology transfer unit.

On the planning and materialization of the change management associated with the ERP project, widespread agreement exists about the convenience of having approached the change management analysis and preparation tasks before the beginning of the implementation project, instead of addressing them during the implementation, mainly in what refers to the integration and standardization of economics management criteria and information among units.
Good Project Scope Management

The interviewed stakeholders said that, in general, the project scope did not vary much along the ERP project. With enough flexibility, the project manager and the steering committee were agreed in their justifications to make some incorporations and changes in the implementation priorities of some functionalities and modules. For example, in the middle of the project, they decided to incorporate the inventory management module, which was not initially planned. Regarding the flexibility in project scope management, some stakeholders complained about the lack of documentation about the decisions taken and lack of justifications in terms of functionality to be implemented, incorporations of new functionality, priority changes, etc., which left these stakeholders with the impression of certain improvisations or decisions being made “on the run.”

Adequate Project Team Composition

With the exception of the software vendor role, the project team covered the rest of areas of interest along the whole project. However, the unanimous opinion, including that of the interviewed implementation consultants, is that the excessive turnover of the implementation consultants complicated the implementation process. With regard to the internal project team, almost all their members continued doing their full working functional activities in parallel with the ERP project. This factor complicated the progress of the project and diminished the satisfaction of team members. And, this factor is given as a justification for the little and poor cooperation of these members with the external implementation consultants.

Following the same line of previous IT/IS department outsourcing decisions, top management decided not to create an internal team of experts in the implemented ERP system; they decided to outsource the future ERP system maintenance. According to some interviewees, the transfer of knowledge to the external consultants should have been made through the IT/IS department personnel specialized in economics management and initially assigned to the project. These people knew extensively the needs of the economics management, the limitations of the existent legacy systems, and some of the limitations of the ERP systems. However, the IT/IS department totally avoided any implementation tasks from the beginning of the project, limiting its intervention to offering services related to the technological infrastructure. Several interviewed stakeholders, among them some from the IT/IS department, related such avoidance to the previous outsourcing of the IT/IS department to a new IT company owned by HEI, with a business model that was at odds with the maintenance of business applications. Some opinions pointed to an additional reason for such an attitude from the IT/IS department: the incorporation of an external project manager for the ERP implementation project, in detriment to the internal leadership from the IT/IS department.

Some interviewees also complained about the insufficient representation of the units in the project team and the steering committee.

Comprehensive Business Reengineering

Before starting the ERP implementation, the steering committee commissioned a project team member to model the existent economics management processes, a task that he completed during the first phase of the implementation project. This factor complicated the progress of the project and diminished the satisfaction of team members. And, this factor is given as a justification for the little and poor cooperation of these members with the external implementation consultants.

Following the same line of previous IT/IS department outsourcing decisions, top management decided not to create an internal team of experts in the implemented ERP system; they decided to outsource the future ERP system maintenance. According to some interviewees,
According to most of the interviewees, after the as-is analysis, both the “to-be” synthesis (description of the redesigned processes) and the “gap-analysis” (description of the differences between the previous processes and the pursued ones) were made on the march for each economics management process approached and were immediately configured on the ERP system. This approach did not allow for the creation of a global transversal model of economics management that would include most of the processes analyzed before their configuration. In general, the ERP system was adapted to the existent processes rather than the other way around.

During the functional analysis phase, no method or modeling tool was used and the final model of economics management processes was not documented.

With regard to the tensions around the analysis and possible process changes, the case of the economics management processes associated with the technology transfer unit deserves special attention. In the steering committee, the people responsible for this unit often showed their strong rejection to changing their processes and their system.

Several interviewees think that in this aspect, a true abyss appeared between the philosophy of operation of the ERP standard for economics management and that of the technology transfer unit. Those practices were very well covered by its legacy system, but they were at odds with the new economics management rules. Interviewees pointed out the following reasons behind the technology transfer unit position: they were used to the autonomy of having their own legacy system; they feared possible loss of personnel, organizational autonomy, and power; and the technology transfer unit rejected the decentralization of part of their processes to the units and to the economics department.

The technology transfer unit case has been the topic of greatest conflict in the whole ERP implementation project in the HEI. After many meetings, debates, and tensions, the ERP implementation in the technology transfer unit has become a big set of bespoke developments around the ERP system, rather than an ERP configuration starting from the ERP standard. On the other hand, the current system has facilitated greater integration between the technology transfer unit and other units.

Adequate Project Sponsor Role

Officially, the figure of ERP project sponsor did not exist. However, there is overall agreement that the then HEI vice president acted as such. One of his crucial functions was to act as mediator. He helped to energize the activities of the team, to define objectives, and to reach small goals, and he contributed decisively to the resolution of the most conflicting situations. As the expert on the management model of the university, because he had lived the whole process of its creation, the vice president helped the people that represented central departments such as the economics department. The units were listened to and they worked together, working out consensus from some initially opposed ideas.

Adequate Project Manager Role

The project manager was hired after the selection of the ERP system, after the selection of the implementation consulting company, and after the creation of the initial project team. According to him, this circumstance already limited his initial performance. His external origin, on one hand, helped him to maintain neutral positions in relation to some organizational conflicts, but was, on the other hand, an initial inconvenience for understanding the organizational culture of the university. There is agreement among the interviewees about the convenience of incorporating somebody external to the organization to manage a project of this type.
A positive aspect of the project manager, commented on by various interviewees, was his open attitude to discussing and to solving conflicts, always trying to follow the legal mechanisms and the appropriate procedures for economics management. The main negative aspect indicated by most of the interviewees was his lack of previous experience in ERP implementations.

Some interviewees noted the lack of monitoring and formal control of the project transmitted by the project manager and the project team in some moments. The project manager and the consulting company did not adopt any implementation methodology, which is also behind the projected uncertainty.

User Involvement and Participation

The implication and wider participation of users refers, according to the interviewees, to the end users of economics department and technology transfer unit. According to the project documentation and the opinion of most of the interviewees, the levels of implication and participation of users from the units was low, probably because the project team and the steering committee did not involve them sufficiently. According to the then vice president, this situation was due to the limitation of budgetary resources that did not allow for the partial liberation of some users from their usual tasks, so that they could be devoted more to the project.

Although both the project team and the steering committee included representative members of the units, some opinions manifested that those people acted more personally than as representatives of the units. Some interviewees indicated that the units were asked for explicit opinions on several occasions, for example in the validation of the definition of the processes, but that the units did not respond to such demands.

Only at the end of the project, in the validation phase, were key users from the units involved and helped key users from the economics department to validate the resulting economics management model. Some economics department users commented that the conjunction of the following facts caused a lot of stress: late implication of the units, lack of time to assimilate the training of the new system, and the coincidence of the end of the year (and the related accounting procedures) with the training and testing activities.

Trust Between Partners

Initially, during the ERP selection and the beginning of the implementation, there was a high level of trust between the HEI, the implementation consulting company, and the ERP vendor branch in Spain. Soon after, a crisis developed in the mutual trust for various reasons.

Regarding the initial relationship between the HEI and the implementation consulting company, most of the interviewees say that the HEI surely trusted excessively in the consulting company. This company at the beginning was sold with the mark or they were hung the medal of being the only consulting company with experience in such a type of ERP implementation in Spain. However, at the beginning, the consulting company omitted that their previous experience consisted in extensive customization and little configuration of the international standard maintained by the ERP vendor.

On the other hand, the ERP vendor in Spain openly referred initially to the HEI as its strategic partner for the HEI sector in Spain, and it committed to facilitating the Spanish standard of all the needed modules by the beginning of the implementation. This did not happen, and, in fact, the ERP vendor several times postponed the delivery of one key module, causing considerable delay in the beginning of the project. Such uncertainty on the delivery of the ERP system caused a lot of insecurity and stress in the implementation consulting company, and in the atmosphere transmitted to the rest of the project team and the steering committee. In some way, the HEI
was aware that the consulting company was also suffering from the lack of commitment and collaboration of the ERP vendor. In fact, the ERP vendor was detaching itself gradually from the project, even ending up with one of its high-level commercial representatives manifesting that the project was no longer considered as “strategic” for his company.

**Consequences**

**Business Process Reengineering (BPR)**

Although BPR started in the design phase, its effective change occurred in the go-live phase. Parallel to the implementation of the ERP system, managers started changing the business processes while explaining them to the organization. The project team admits that this should have been done before the implementation, because some processes were obsolete or inadequate. Currently, during the postimplementation period and using the knowledge that the project team acquired, they are continuously improving processes by extending functionality through the current discovery of functionality in the ERP system. The ERP system brought in new business processes and is helping in the reengineering of the old ones. As the project manager mentioned, the HEI does not have any distinctive economics management processes. Therefore, they would not have lost competitiveness by adopting the ERP best practices. On the contrary, the adoption of practices that the ERP provides helped to reengineer the existent business processes and simplified implementation and future maintenance efforts.

**Change of Mentality**

One of the most interesting consequences of this ERP implementation was the induced change of mentality. At the beginning, most of the users disagreed with the implementation of the system, but their attitudes reversed progressively after the go-live phase. Nowadays, they recognize that the ERP system is useful and that it has helped to improve the HEI processes. They think that more business analysis is still necessary, with managers demanding more people to improve the work. Some users were moved from their old functions, while others had intensive training. Now, only a few employees are still not using the system. Some of them are managers of small units that do not have enough economics management documents to justify one person being allocated to economics management tasks. The solution has been to create a special group in the economics department that manages all the economics management documents from the small units. Now, users are not afraid of changes. They are aware that the system is about continuously improving processes.

**Lack of Internal ERP Knowledge**

Following the same line of previous IT/IS department outsourcing decisions, top management decided not to create an internal team of experts in the implemented ERP system; they decided instead to outsource the future ERP system maintenance. Given the nature of the implementation carried out, this strategy is questioned by many stakeholders who consider the current situation as one of excessive dependence with regard to the implementation consulting company. HEI managers now know that they are totally dependent on the consulting company, and they are trying to change this in the near future.

**CONCLUSIONS AND IMPLICATIONS FOR FURTHER RESEARCH**

In this chapter, we attempted to provide a view of the ERP implementation process in an HEI. To the best of our knowledge, we present the first grounded theory model for ERP implementa-
Understanding an ERP System Implementation in a Higher Education Institution

tions in HEI. We think that the main conclusion of this study is the evidence that organizational and cultural issues have a strong impact in ERP implementation projects. In the particular case addressed in this study, if those issues had been taken into account during the planning phase, it is very likely that some problems would have been avoided or at least mitigated.

In our case study, the findings suggest that organizational culture and structure may have an influence on the management of critical success factors. These could be helpful in explaining some ERP project issues and, therefore, defining a strategy for ERP implementation. Research on critical success factor identification does not explain why some factors are more relevant in some organizations than others. Cultural issues can be responsible for this. The influence of culture has mainly focused on the IS solution and how it fits with the organizational culture or the cultural changes.

We believe that our study of the ERP implementation process is a valuable contribution to the literature, yet the generalizability of our propositions is limited due to the scope of the sample, only an in-depth case study. We would like to emphasize that this was an in-depth case study combined with grounded theory method that took more than a year of documentation analysis, interviewing, and data coding and analysis.

REFERENCES


Chapter 6.11
Ontologies for Scalable Services-Based Ubiquitous Computing

Daniel Oberle
SAP Research, CEC Karlsruhe, Germany

Christof Bornhövd
SAP Research, Research Center Palo Alto, USA

Michael Altenhofen
SAP Research, CEC Karlsruhe, Germany

ABSTRACT

This chapter discusses scalability problems and solutions to services-based ubiquitous computing applications in real time enterprises. The scalability problems are (1) identifying relevant services for deployment, (2) verifying a composition by a logical rule framework, and (3) enabling the mapping of required services to the “best” available device. We argue that ontologies can help to counter these challenges. Subsequently, we provide a detailed introduction to ontologies. We focus on the ontology languages emerging from the corresponding W3C Semantic Web activity. The W3C recommendations have a high impact on future tools and the interoperability of ontology-based applications. We contrast the pros and cons of ontologies at a general level and demonstrate the benefits and challenges in our concrete smart items middleware.

INTRODUCTION

Ubiquitous computing for real time enterprises calls for novel approaches to distributed applications since both the economic and technical scale of these applications will increase dramatically. Regarding the economic scale, applications grow beyond enterprise boundaries and potentially involve frequently changing, partly unknown participating entities. Therefore, much more open
Ontologies for Scalable Services-Based Ubiquitous Computing

approaches are desired. In the world of enterprise computing, service-oriented architectures and Web services are considered important steps on the road to economic scalability as discussed in the chapter “Ubiquitous Services and Business Processes.”

The concepts and technologies of service-oriented architectures can be fruitfully applied to ubiquitous computing for real time enterprises to counter the economic scale. However, the technical scale of services-based ubiquitous computing, which concerns the number of nodes and processes involved, still remains a challenge. We discuss the use of ontologies as a possible means of countering the scalability problem, and evaluate their promises and limitations for services-based ubiquitous computing applications.

Services-based ubiquitous computing applications are highly distributed applications that run in the form of cooperating services on a variety of possibly heterogeneous devices. The devices run services that can be combined, that is, “composed,” into more complex services or applications. Such a services-based approach to the development of ubiquitous computing supports the distribution of functionality across the set of available devices, enables better reusability of components in new or different applications, and the division of the overall functionality into independent services with clearly defined interfaces that can be developed and tested separately.

However, the capability to decompose business processes into individual services and to deploy them on different smart devices poses new technical challenges. In particular, the services to be executed on the devices need to be modeled and described for identification and selection, mapped to appropriate smart devices, remotely deployed, configured, and monitored. The reason for such additional tasks is the heterogeneity of the underlying hardware platforms in terms of different communication protocols and technologies.

To facilitate such tasks, a service composition model tailored to smart device interaction is required. The service composition model has to enable the explicit modeling of the heterogeneities of different device platforms and has to support the identification, deployment, and composition of services for smart devices. While the identification, deployment, and composition are fairly simple with a small number of services and devices, the challenge of coping with such tasks increases with the number of devices and services.

In the remainder of this chapter, we show how ontologies may counter the fundamental scalability challenges of: (1) identifying relevant services for deployment, (2) verifying a composition by a logical rule framework, and (3) enabling the mapping of required services to the “best” available devices.

We begin by detailing the scalability challenges in a service-oriented smart items middleware. The chapter continues by introducing the reader to ontologies. We focus on the recent W3C recommendations RDF(S) and OWL that, for the first time, provide standard languages for ontology specification. These recommendations have a high impact on future tools and the interoperability of ontology-based applications. We proceed by contrasting the pros and cons of ontologies on a general level. Finally, we sketch how ontologies are used to counter challenges (1)-(3) in a service-oriented smart items middleware.

CHALLENGES IN SERVICE-ORIENTED SMART ITEMS MIDDLEWARE

As mentioned in the introduction, the technical scale of services-based ubiquitous computing applications struggles with the heterogeneity of the underlying smart devices, and the inherent complexity of a system that requires automatic or semi-automatic monitoring and deployment of services. In this section, we refer to the service-oriented smart items middleware presented in the
chapter “Smart Items in Real Time Enterprises” and detail where such challenges occur. We argue that ontologies should be used to describe services and devices. Ontologies provide the means to identify available services, verify given service compositions, and guide the deployment of services.

**Smart Items Middleware**

The architectural overview of smart items middleware in the chapter “Smart Items in Real Time Enterprises” introduces five logical system layers. The technical scalability challenges occur in both the Device Layer and Device Level Service Layer. We would like to recap these layers in the following paragraphs.

The **Device Layer** comprises the actual smart item devices, the communication between them, and the presentation of the available hardware services to the next higher layer. Available devices usually provide different communication means and operating systems. For example, in the case of RFID, the reader always initiates the communication and the tags cannot directly communicate with each other. In case of wireless sensor networks, simple services can be pushed and executed on them and tasks can be accomplished in cooperation, since the sensor devices have integrated processing power and are able to use peer-to-peer communication.

The **Device Level Service Layer** manages the deployable services used by the device layer. It contains a service repository that stores a service description and one or more service executables for each service. Compound services rely solely on other services to fulfill their task and have a service composition description that is stored in the service repository. Atomic services, on the other hand, are associated with directly executable code. Since a service may be deployable on different platforms, an atomic service may have more than one service executable—one for each device platform.

**Running Example**

In order to illustrate the challenges in the smart items middleware, we first describe a running example (see Figure 1). We assume we are in a warehouse, where goods are tagged with RFID chips and shelves are equipped with RFID readers. The RFID readers feature both Ethernet and encrypted WLAN communication. Further, we assume that temperature sensors and gateway nodes are deployed at strategic locations within the warehouse. The gateway devices are linked to the RFID readers in the smart shelves via Ethernet cables and are equipped with antennas for wireless communication with the temperature sensors (e.g., via 433MHz) and for nonencrypted WLAN (about 2.4 GHz). Every warehouse employee is equipped with a PDA. The PDAs are integrated into a WLAN and provide an adapter for communication with the sensor devices.

Warehouse employees should be informed via their PDAs whether perishable goods are on shelves where the measured temperature exceeds a predefined threshold. The implementation and deployment of the corresponding services in the middleware can take place in the following three main steps. In the first step, a developer chooses

![Figure 1. Running example: Goods, RFID, and temperature sensors in a warehouse](image-url)
the services from the middleware service repository required for our application. In the second step, the developer implements the new application by composing services and device types. Finally, the services of the verified composition plan are mapped to the appropriate devices. Each of the three steps is afflicted with fundamental scalability challenges.

**Fundamental Scalability Challenges**

As the number of services and devices increases, identifying required services, verifying service compositions, and mapping services to devices becomes correspondingly more complex for the system developer. In the following section, we discuss three fundamental scalability challenges of the smart items middleware, which occur independently of any application. A technical scalability challenge is one that increases with the number of services and devices. Tool support is required because the developer can no longer handle the challenges manually.

**Challenge 1: Service Identification**

The descriptions of services available for deployment in the smart items middleware are stored and maintained in the middleware’s service repository. At system deployment or during service composition, relevant services must be selected from the repository. The challenge here is to enable selection based on the semantics or behavior of the available services, rather than on the available interfaces or service signatures.

In our running example, the developer should be able to conveniently search for a “temperature sensing” service rather than having to conduct the tedious task of checking and examining all the services’ interfaces.

**Challenge 2: Verification of Service Compositions**

Services available in the service repository can be composed into new, more powerful, services or applications. However, when composing services at the level of smart devices, heterogeneities of the underlying hardware platforms need to be addressed (Ziekow, Avanes, & Bornhövd, 2006). Heterogeneities concern different communication protocols and technologies as well as different programming models. An abstraction to a common, homogeneous service platform is not possible due to the limited computing capabilities of typical smart devices. The challenge here is to verify a composition plan with respect to the heterogeneities of the underlying hardware.

In our running example, the gateway devices can be used with a service that informs all PDAs about perishable goods that are too warm. As an alternative, the PDAs could also directly receive and evaluate the sensor values, and could request the required RFID data. However, in the setting described, there is no communication channel shared by RFID readers and PDAs. Composition plans for both situations can be specified, but only the first one can be implemented.

**Challenge 3: Intelligent Service-to-Device Mapping**

In addition to manual deployment using dedicated deployment tools, the system should be able to automatically map required services onto the most suitable available devices, or relocate services from one device to another based on changing system conditions. Such automatic deployment of services is essential to manage the increasing complexity of a distributed smart items solution (Ziekow et al., 2006).

In our running example, it would be worthwhile to define a deployment constraint that the
Ontologies for Scalable Services-Based Ubiquitous Computing

temperature filter service gets deployed on only 30% of the nodes, and that these nodes also have to have a certain amount of battery power left. This constraint imposes the need to remap services, if some sensor nodes are running out of power.

In the remainder of this chapter we argue that ontologies can help to counter these challenges. Therefore, the section “Ontologies and Semantic Web Specifications” continues with a general introduction to ontologies. We focus on the ontology languages emerging from the corresponding W3C Semantic Web Activity. The W3C recommendations have a high impact on future tools and the interoperability of ontology-based applications.

What is a Conceptualization?

A conceptualization is a representation of a relevant domain in a more or less formalized model expressed via cognitive modeling primitives. Modeling primitives are cognitive, when they resemble the way humans think and conceive of the world. Ontologies usually provide classes, relations between classes, and instances as cognitive modeling primitives. The backbone of every conceptualization, and thus, each ontology is a hierarchy of classes (also called a taxonomy). This means that sub- and superclass relations are very prominent in ontologies and that special modeling primitives are devoted to them.

Let us model the domain of our running example to clarify this notion. We start by identifying relevant classes in the warehouse domain and putting them into a taxonomy. In the section “Running Example” we learn about Ethernet cables, antennae, RFID chips, sensors, gateways, and so forth. As a first categorization we introduce the classes Specification, Communication Means, and Node. The three classes become direct subclasses of Entity (the most general class). In turn, we classify Antenna and Cable underneath Communication Means, and Sensor, Gateway, Reader, and PDA underneath Node. Furthermore, we introduce the relation describes between Specification and Communication Means as well as accesses between Node and Communication Means, as shown in Figure 2.

With a set of instances, we are able to capture a specific setting in our domain. Figure 3 depicts a specific situation where we can find specific Nodes, viz., Gateway#1, RFIDReader#1, and PDA#1. All of them are linked to their respective Communication Means via the accesses relation. In turn, concrete Specifications, that is, WLAN, WLANencrypted, Ethernet, describe the Communication Means. The reader might wonder, however, where to start and when to stop model-
ing. This leads us to the following requirements for a conceptualization:

Relevance

The conceptualization must represent only relevant classes, relations, and instances. Relevance depends on what we want to achieve with the conceptualization. In our case, we would like to counter the three challenges mentioned in the previous section. For this purpose, it is unnecessary to represent the warehouse as a class, for instance. With respect to these criteria, all classes and relations depicted in Figure 2 are relevant.

Completeness

The conceptualization must represent all relevant classes, relations, and instances. In the running example, Specifications, Communication Means,
and Nodes are all relevant, but will not suffice to address the three challenges. Thus, the conceptualization depicted in Figure 2 is not complete. We have skipped many details for the sake of brevity. The interested reader may refer to Oracle (2006) and Spieß, Bornhövd, Haller, Lin, and Schaper (2007) for the details.

Correctness

The conceptualization must reflect the domain in a consistent way. That means each class, relation, and instance must have counterparts in the domain. This requirement holds for the conceptualization shown in Figure 2.

This way of modeling a domain is very similar to UML class diagrams (Booch, Jacobson, & Rumbaugh, 1998) in object-orientation or entity relationship models (Chen, 1976) in database design. However, ontologies require an explicit, logic-based representation of the domain, as explained in the next section.

What is an Explicit Specification?

The definition of an ontology as given by Gruber (1995) requires an explicit specification of the conceptualization. This means that ontologies must be specified in a formal, logic-based language. Such languages typically provide a wealth of modeling primitives and can avoid ambiguities bound to graphical notations. In addition, the underlying logic is a prerequisite for automatic reasoning, that is, drawing inferences on the basis of the conceptualization. This is in stark contrast to purely graphical modeling techniques. These can be seen as conceptualizations as well, yet are only implicitly specified by graphical representations.

Until recently, there has been a plethora of logic-based ontology languages and corresponding tool suites in research and academia which do not allow any interoperability. However, the W3C’s Semantic Web Activity (W3C, 2001) has gained momentum and specified several recommendations (“standards,” in the W3C jargon). The recommendations (in particular RDF and OWL) are a significant step towards tool interoperability and will thus have a high impact in the future.

According to Tim Berners-Lee, the semantic Web augments the current WWW by giving information a well-defined meaning, better enabling computers and people to work in cooperation (Berners-Lee, Hendler, & Lassila, 2001). This is done by adding machine-understandable content to Web resources. The result of this process is metadata, usually described as data about data. Descriptions such as this acquire semantics by referring to an ontology. The semantic Web’s vision is that once all the layers shown in Figure 4 are in place, we will have an environment in which we can place trust that the data we are seeing, the deductions we are making, and the claims we are receiving have some value. The goal is to make the user’s life easier through the aggregation and creation of new, trusted information from the Web, to enhance search functionality, and, in the ideal case, to infer additional knowledge.

Figure 4 depicts the original “layer cake” proposed by Tim Berners-Lee (Berners-Lee et al., 2001)
The role of Unicode, URIs, XML, and Namespaces is limited to that of a syntax carrier for data exchange. An XML Schema defines simple data types such as string, date, or integer. The standardization process has currently reached OWL; the remaining layers are not yet fully specified. Hence, many changes to the original layer cake are possible and quite likely. In the following section, we discuss the layers one by one. In doing so, we focus on the two most stable recommendations so far, namely RDF and OWL.

**RDF(S)**

The resource description framework (RDF) (Lassila & Swick, 1999) can be used to make simple assertions about Web resources or any other entity that can be named. A simple assertion in RDF is called a statement and consists of a triple (subject, predicate, object). Subjects, predicates, and objects are URIs; objects may also take a literal value (e.g., a string). For instance, a statement (example:WLAN, rdf:label, “IEEE 802.11 WLAN Specification”) associates a name with the entity identified by the example:WLAN URI.

RDF Schema (RDFS) extends RDF with additional modeling primitives (Lassila & Swick, 1999), viz., classes and relations (called properties in RDFS) that can be specified, put into hierarchies, and interlinked to specify simple ontologies. For example, the classes, relations, and instances shown in Figures 2 and 3 can be specified by a set of RDF(S) statements. Below, we list the statements that define the Specification class, the describes relation, and the example:WLAN instance.

(example:Specification, rdf:type, rdfs:Class)  
(example:Specification, rdfs:subClassOf, example:Entity)  
(example:describes, rdf:type, rdfs:Property)  
(example:describes, rdfs:domain, example:Specification)

**OWL**

RDF(S) has one SQL-like query language, which is called SPARQL (Prud’hommeaux & Seaborne, 2006). However, the lack of a formal logical underpinning of RDF(S) prevents automatic reasoning tasks. In addition, the RDF(S) modeling primitives are limited with respect to expressiveness. Therefore, the W3C has come up with a dedicated set of logic-based languages called OWL (Web ontology language) (McGuinness & Harmelen, 2004). This set of languages allows the explicit specification of conceptualizations by logical theories, making intelligent reasoning tasks possible. For instance, using OWL, relations can be specified as transitive, and the transitive closure of the relation can be inferred automatically. OWL consists of:

**OWL Lite**

OWL Lite is the simplest variant of OWL. Its modeling primitives are a strict subset of the OWL DL modeling primitives. The modeling primitives were chosen using an 80:20 rule: OWL Lite contains the modeling primitives which are most often required. Furthermore, OWL Lite’s reduced expressiveness leads to more efficient reasoners.

**OWL DL**

OWL DL is the most prominent variant of OWL. It is based on an older variant of Description Logic.
Ontologies for Scalable Services-Based Ubiquitous Computing

(Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2003) which essentially has been made Web-compliant. Description Logics are a family of knowledge representation languages which can be used to represent the terminological knowledge of an application domain in a structured and formally well-understood way. Web compliance means that URIs are used for identification, XML schema datatypes are introduced, and there is an RDF serialization for representing OWL DL ontologies. Below, we list the modeling primitives that exceed the expressiveness of RDF(S)². The additional modeling primitives can be leveraged to specify our conceptualization more precisely:

- **Transitive Relations:** If a relation \( P \) is specified as transitive, then \( P(x,y) \) and \( P(y,z) \) implies \( P(x,z) \).
- **Symmetric Relations:** If a relation \( P \) is tagged as symmetric, then \( P(x,y) \) holds if \( P(y,x) \).
- **Functional Relations:** If a relation \( P \) is tagged as functional, then \( P(x,y) \) and \( P(x,z) \) implies \( y = z \).
- **Inverse Relations:** If a relation \( P_1 \) is tagged as the inverse of \( P_2 \), then \( P_1(x,y) \) holds if \( P_2(y,x) \).
- **Inverse Functional Relations:** If a relation \( P \) is tagged as inverse functional, then \( P(y,x) \) and \( P(z,x) \) implies \( y = z \).
- **allValuesFrom:** The allValuesFrom modeling primitive restricts the range of a relation to a (complex) class. In our running example, we would restrict accesses to Communication Means, for instance.
- **someValuesFrom:** The difference between this and allValuesFrom is that there must exist at least one instance in the range of the specified relation.
- **Cardinality:** The cardinality primitive permits the specification of exactly the number of elements in a relation.
- **hasValue:** The hasValue primitive allows us to specify classes based on the existence of particular relation values.
- **equivalentClass:** The equivalentClass primitive is used to indicate that two classes have precisely the same instances.
- **equivalentRelation:** In order to tie relations together in a similar fashion, we use equivalentRelation.
- **Identity between Instances:** This primitive is similar to that for classes, but declares two instances to be identical.
- **Different Instances:** This primitive provides the opposite effect from the previous.
- **Complex Class Definitions:** OWL DL provides additional primitives for forming complex classes. The basic set operations, namely union, intersection, and complement, are provided. Additionally, classes can be defined by enumeration, and it is also possible to assert the disjointness of classes.

**OWL Full**

As the name suggests, OWL Full supersedes the expressiveness of OWL DL. For example, the distinction between classes and instances is no longer strict. Instances can also be classes simultaneously, and can have instances themselves. In essence, OWL Full equals first-order logic (also called predicate logic), which is known to be undecidable. This means that sound and complete reasoners cannot be constructed for this logic, which makes it very impracticable.

The explicit specification, that is, logical representation, of a conceptualization is a prerequisite for automatic reasoning tasks. Reasoners (also known as inference engines) enable such tasks, as they are implementations of logic calculi. Examples of OWL reasoners are KAON2 (Motik, Sattler, & Studer, 2004), FaCT (Horrocks, 1998) and Racer (Haarslev & Möller, 2001). Next, we give a partial overview of typical OWL DL reasoning tasks:
Class satisfiability: A class is satisfiable with respect to an ontology if the ontology can be interpreted such that the extension of the class is nonempty, that is, the class can potentially have an instance.

Ontology consistency: An ontology is consistent if all of its classes are satisfiable.

Class subsumption: A class X subsumes a class Y with respect to an ontology if any instance of Y is also an instance of X, no matter how the ontology is interpreted. A pairwise subsumption test of all classes in the ontology obtains an automatic classification of the taxonomy.

Class equivalence: A concept X is equivalent to a class Y with respect to an ontology if X and Y subsume each other.

Class disjointness: Two classes are disjoint with respect to an ontology if they do not have a common instance.

Instance retrieval: Instance retrieval specifies which instances in the ontology are subsumed by a query.

Instance realization: Instance realization specifies to which classes an instance belongs.

Rules

The Rule Layer provides an interoperable language for describing the sets of deductions one can make from a collection of data, that is, how, given a ontology-based information base, one can derive new information from existing data. At the time of writing this chapter, SWRL, the Semantic Web Rule Language (Horrocks, Patel-Schneider, Boley, Tabet, Grosof, & Dean, 2004), F-Logic (Kifer, Lausen, & Wu, 1995), and Business Rule approaches (Morgan, 2002), are in discussions for standardization within the context of the W3C Rule Interchange Format (W3C, 2005).

In our running example, we can specify a rule which automatically infers whether there is a communication channel between two nodes. This is the case when two nodes access a communication means with identical specification. The corresponding rule is depicted next, and can be leveraged for countering challenge (2) as detailed in the section “Ontologies for Verification of Service Composition.” channel(Gateway#1,PDA#1) would be inferred because both nodes access the same Ethernet cable. A counterexample would be channel(RFIDreader#1,PDA#1) because the nodes access antennae with incompatible WLAN specifications.

\[
\text{channel}(x,y) \leftarrow \text{Node}(x) \land \text{Node}(y) \land \text{accesses}(x,a) \land \text{accesses}(y,b) \land \text{Specification}(z) \land \text{describes}(z,a) \land \text{describes}(z,b)
\]

Proof, Trust, and Digital Signatures

The Proof language will provide a way of describing the steps taken to derive a conclusion from the facts. Proofs can then be passed around and verified, providing shortcuts to new facts in the system without requiring each node to make the deductions on its own. The vision of the Trust layer on top is to create an environment in which we can place trust on the data we are seeing, the deductions we are making, and the claims we are receiving. For reasoning to be able to take trust into account, the common logical model requires extension to include the keys with which assertions have been digitally signed. The Proof and Trust layers, as well as the Digital Signatures, are not yet specified.

BENEFITS OF ONTOLOGIES

In the previous section, we introduced ontologies and focused on recent W3C recommendations emerging from the Semantic Web Activity. The latter are likely to have a high impact on future tools and the interoperability of ontology-based applications. In this section, we discuss the general benefits of ontologies. We give concrete examples
of the benefits in the section “Ontologies for Improving the Smart Items Middleware”, where we detail how ontologies can be used to counter the three challenges introduced in the section “Fundamental Scalability Challenges”.

**Benefits through Conceptual Modeling (Conceptualization)**

The first set of benefits concentrates on conceptual modeling. All the advantages are conceivable at first sight. However, they are hard—or sometimes even impossible—to measure.

**Reuse, Flexibility, and Maintainability**

With modeling techniques, such as UML, ERM, or ontologies, we make the conceptual model underlying a specific application explicit. In our running example, we analyze the domain of smart devices, identify relevant classes and their relations, and represent this knowledge explicitly. Thus, the ontology can be reused in other applications in the same domain, and can be used to facilitate communication with domain experts. The alternative would be to hard-code domain knowledge, resulting in an inflexible application with high maintenance costs (every change of the domain knowledge would require code adaptation and recompilation).

**Reuse of Best Practice for Conceptual Modeling**

Recent developments allow us to talk about the quality of ontologies as well as ontology design patterns. In addition to purely structural approaches, for example, the normalization of relational data models, the ontology design patterns consider the conceptual level (Gangemi, 2005). For example, a simple participation pattern (including objects taking part in events) emerges in domain ontologies as diverse as enterprise models, legal norms, software management, biochemical pathways, and fishery techniques. Ontology design patterns are useful to acquire, develop, and refine an ontology.

**Standardization**

We already learned in the section “Section Ontologies and Semantic Web Specifications” that the standardized ontology languages of the W3C are a significant step towards tool interoperability and will thus have a high impact in the future. So far, there have been a plethora of ontology languages in research and academia without any interoperability. In addition, the Object Modeling Group (OMG) is currently developing an Ontology Definition Metamodel (IBM, Sandpiper, 2006). Once established, ODM will allow using UML tools to specify OWL ontologies as well.

**Benefits from Explicit Specification**

The second set of benefits focuses on the explicit specification, that is, advantages that are a consequence of the logic-based representation. As with the aforementioned benefits, the following ones are also hard to substantiate in numbers.

**Unambiguous, Formal Representation of Shared Meanings of Terms**

The advantage of ontologies over other conceptual modeling techniques is the decreased ambiguity of the modeling. This is due to the increased formality of the logic-based representation. For instance, relations can be defined as symmetric, transitive, or inverse. A reasoner can leverage the formal representation to infer additional knowledge, perform consistency checks, and so forth. A user of the ontology can leverage the increased formality to avoid ambiguity. Depending on the expressiveness of the ontology language, one might add further rules or integrity constraints. While
something similar can be done with approaches such as OCL (the Object Constraint Language in UML), ontologies demand agreement at the modeling level. All this results in a formal, machine executable description of the shared meaning of terms and relations.

Reasoning Saves Development Efforts

We discussed in “Ontologies and Semantic Web Specifications” that ontology languages are based on logics, possibly with a rule mechanism on top. Reasoners let us leverage the advantages of the logics and offer powerful functionality, such as instance retrieval (all instances of a class $x$), subsumption checking (is a class $x$ subsumed by a class $y$?), automatic classification of an ontology, consistency checking, and potentially a rule mechanism. This means that such functionality does not have to be developed from scratch.

Many problems can be reduced to such functionality. For example, the matching of services can be reduced to the subsumption problem. A detailed example follows in the section “Ontologies for Improving the Smart Items Middleware,” where we apply rules to realize a verification mechanism. In essence, we can say that ontological modeling and the associated reasoning capabilities can be used to save development efforts.

Standardization

So far, there has not only been a plethora of ontology languages in research and academia, but also a multitude of existing ontology editors, stores, and reasoners. Hence, standardization also comes in handy for interoperability between such tools.

LIMITATIONS OF ONTOLOGIES

Despite all the benefits discussed in the previous section, every ontology-based application encounters limitations. In the following section, we proceed by giving a general explanation of each limitation. We give concrete examples of the limitations in the section “Ontologies for Improving the Smart Items Middleware.”

Theoretical Limitations

The first set of limitations is of a theoretical nature. Among them is the fundamental trade-off between expressiveness and efficiency, as well as the widely known frame problem.

Expressiveness vs. Efficiency

Representation languages for ontologies encounter a trade-off between expressiveness and efficiency. The search for expressiveness is an elaborate process of finding consensus on what modeling primitives might be required most urgently by potential users. Requirements for more modeling primitives, that is, expressiveness, must be traded off against efficiency or even decidability of the corresponding reasoner. In essence, we can say that the more modeling primitives the language provides, the less efficient the respective reasoners are. The decision for the right language has to be taken on a case-by-case basis: if expressiveness is important, use a language with many modeling primitives but with a less efficient reasoner. If fast run-time reasoning is important, use a less expressive language.

The most recent example of this trade-off is shown by the three increasingly powerful variants of OWL, viz., OWL Lite, OWL DL, and OWL Full. OWL Lite and OWL DL are languages from the family of description logics (DL) and, thus, strict subsets of first-order logic. If chosen carefully, as are OWL Lite and OWL DL, they are decidable, which means that sound and complete reasoners can be constructed. OWL Full equals full first-order logic which is known to be undecidable, that is, no sound and complete reasoner exists for this language.
The Frame Problem

Ontologies are ideally suited to categorize the entities in a relevant domain and to capture their interrelationships, as well as specific situations, via instances. User groups can agree on the ontology with the ultimate goal of information integration and mutual understanding. However, ontologies exhibit fundamental limitations when describing the behavior of entities. As an example, consider the preconditions, postconditions, and the functionality of a service. All these aspects are required for semantic service discovery, yet their representation via ontologies is severely limited. The reason is to be found in the logic-based representation and is called the frame problem.

Put succinctly, the frame problem in its narrow, technical form is this: Using mathematical logic, how is it possible to write formulae that describe the effects of actions without having to write a large number of accompanying formulae that describe the obvious noneffects of those actions? This means that in standard first-order logic, one would need to describe the properties that change as the result of an action as well as the properties that do not change (Baader, 1999). The challenge is to find a way to capture the noneffects of actions in formal logic. What we need is some way of declaring the general rule-of-thumb that an action can be assumed not to change a given property of a situation unless there is evidence to the contrary. The main obstacle to doing this is the monotonicity of classical logic. In classical logic, the set of conclusions that can be drawn from a set of formulae always increases with the addition of further formulae. This makes it impossible to express a rule that has an open-ended set of exceptions. The solution to the frame problem requires special logic languages which are, however, not aligned with common ontology languages, for example (Levesque, Pirri, & Reiter, 1998).

Practical Limitations

Unlike the theoretical limitations, which are rooted in the fundamentals of ontologies, the limitations discussed in this section typically occur when ontologies are applied in practice.

Development vs. Modeling Efforts

In the section “Benefits from Explicit Specification,” we learned that reasoning could be used to reduce the development efforts. Reasoners provide us with powerful functionality, such as consistency checks, automatic classification of an ontology, rules, and so forth. A specific problem might be reduced to the services offered by the reasoner. Therefore, one is able to avoid developing a solution from scratch.

The drawback here is that we have to invest modeling efforts to apply the reasoner. First of all, developers have to be skilled enough to build an ontology and rules. While building the actual ontology is a one-off effort, every relevant entity has to be modeled via the ontology as well. This is a linear effort based on the number of entities. Finally, we cannot expect the ontology to remain stable over time. Changes will be necessary, and will lead to maintenance efforts.

In our example, we have to come up with the ontology depicted in Figure 2 and extensions for specific domains. While this is a one-off effort, it is not to be underestimated. The ontology has to be built in such a way that the rules can be efficiently applied. This requires a lot of expertise from the ontology engineer. Furthermore, all entities in our example have to be classified and modeled via the ontology. As shown in the section “Ontologies and Semantic Web Specifications,” the WLAN instance has to be classified as an instance of the Specification class and has to be related to the ontology shown in Figure 3. As mentioned before, this is a linear effort based on
the number of entities we have to model. In addition, the rules must be modeled, which requires detailed domain expertise.

In conclusion, there is a trade-off between development and modeling efforts. On the one hand, we save development efforts by reducing specific problems to the functionality provided by reasoning. On the other hand, we have to invest modeling efforts in order to model the ontology and relevant entities. Modeling is a prerequisite for using the reasoner. Identifying the sweet spot between both efforts is a difficult undertaking, because it is hard to measure. Therefore, the trade-off must be made on a case-by-case basis.

Modeling Scope

Ontologies are afflicted with several other problematic aspects, which we subsume under the term “modeling scope”. First, ontologies formalize the shared understanding of a user group to enable information integration. However, different user groups will use different ontologies: we cannot expect the usage of an all-encompassing ontology. Manual reconciliation or sophisticated ontology mapping algorithms will be required. (Ehrig, 2006) discusses several different approaches to ontology mapping. Most of them are probabilistic and require manual revision. Another approach is to use a common foundational ontology to align the ontologies (Guarino, 1998). The foundational ontology can be seen as a generic modeling basis which is used for several domain-specific ontologies. It is expected that having a common basis facilitates mapping between domain-specific terms. However, this has not been proven so far.

Second, ontologies model a specific domain. There have been attempts for an all-encompassing ontology, but they have failed, typically because of the lack of agreement and because of the rapidly changing and expanding universe of discourse. One of the most prominent examples of the past has been the CYC project (Guha & Lenat, 1990). In our example, we also encounter expansions of the universe of discourse, that is, the ontology must account for future technologies.

ONTOLOGIES FOR IMPROVING THE SMART ITEMS MIDDLEWARE

In this section, we leverage the advantages of ontologies to counter the challenges (1) - (3) presented in the section “Fundamental Scalability Challenges.” Of the three challenges, we sketch solutions to (1) and (3) and discuss (2) in more detail. Furthermore, we highlight which of the benefits and limitations apply.

Ontologies for Service Identification

Descriptions of services available for deployment in the smart items middleware are stored and maintained in the service repository. At system deployment or during service composition, relevant services must be selected from the repository. This selection usually takes place based on the available interfaces or service signatures, rather than on the semantics or behavior of the offered services. Ontologies can be used as the common vocabulary for the semantic description of services in the repository. They provide the formal basis for the mapping of a service request to available services (cf. benefits unambiguous, formal representation of shared meanings of terms as well as reuse, flexibility, and maintainability). However, such semantic service descriptions do not include the specification of the operation performed, for example, provision of temperature readings or the filtering of incoming sensor readings, as well as pre and post conditions of a service (cf. limitation frame problem).

In our running example, a developer chooses the services from the service repository required for our application. A request for a “temperature sensing” service gets translated by the repository into a request based on the common vocabulary used to describe the semantic inputs and outputs
of the available services. Through a couple of refinement steps, the system enables the developer to identify the required services. In a similar way, the developer identifies the desired device types based on their profile and availability.

**Ontologies for Verification of Service Composition**

Ontologies provide the means to conceptually model the heterogeneities of different device platforms. Software, platform, and device components must be enriched by additional information about the communication technologies they apply. The biggest challenge in modeling is the heterogeneity of communication aspects of different device types that must be represented. Required attributes may range from shapes of plugs to frequencies of radio signals. For example, to model applications for a system using different types of data cables, details about plug shapes and their specifications must be represented in the ontology. Also, in order to make the ontology adaptable to details of every application domain and to enable extensions for future technologies, it must be extensible.

The problem with this flexibility is in finding verification algorithms that work in spite of changes in the structure of the ontology. This problem can be solved by using the rather generic ontology depicted in Figure 2. This ontology serves as the basis on which more specific domain ontologies can be defined.

Combining the modeled information, the required and provided technologies for cross device interaction can be automatically derived and used to verify composite applications by a reasoner. The verification is achieved by compatibility rules for composition validation defined only on the basis of classes and relations in the ontology. Consequently, these rules apply to all descriptions that are based on this ontology. We have already seen an exemplary rule in the section “What is an Explicit Specification?” Therefore, it is ensured that the descriptions can be adapted to the required level of granularity and can be extended in the future. For example, new protocol standards to access upcoming generations of RFID tags must be included in the ontology once they become available. As long as the generic ontology is respected, no changes in the verification algorithms are needed. With this solution, we leverage the benefits of reasoning saves development efforts as we do not have to realize the verification from scratch. The drawback of this approach is that we have to expend efforts for modeling the ontology and rules, which can only be done manually (cf. limitation development vs. modeling efforts).

**A Prototypical Implementation**

Ziekow et al. (2006) provides a first prototype of a verification tool that can be used to compose and verify service compositions for smart device applications. As input, it uses an ontology together with verification rules. In addition, the tool requires descriptions of compositions (composition plan). These are models of functional components, for example, services or devices, and a specification of how these components are combined into a more complex service or application.

The verification tool includes a reasoner that checks the rules against the given composition plan. In this way, the correctness of the modeled composite service or application is determined. The tool returns a report about the check results. This can include errors and warnings or may just indicate a correct composition. In case of an error, information about the cause of a failure is provided to support the correction of the encountered problems.

During the verification process, intermediate results of composition steps are created. These are models of composite components resulting from the respective composition step. These models are also part of the verifier’s output and describe already-verified combinations of components, which can be stored and reused for future compositions.
In this way, the tool supports the development and utilization of composite services or applications for collaborative smart devices. Detailed knowledge about the low-level communication issues is modeled once and automatically used by the tool. Another system component or a human can use the described tool to check for correctness of a composition, leaving the burden of considering details of the communication layer to the tool.

Revisiting our Running Example

In our running example, the developer might want to reduce energy-intensive communication. For deployment on the nodes, a service is chosen which identifies and filters temperature readings to the gateway when they exceed a specified threshold. For the gateways, a service is selected that issues read-requests to the RFID readers if more than 15% of the sensors report temperature readings above the threshold. Based on the RFID data, it can be determined if perishable products are on the shelves. Another service on the gateways informs all PDAs about perishable goods that are too warm. The specified composition of services and device types can now be validated based on the service description and device profiles.

From a software perspective, an alternative implementation without using gateway devices would have been possible. The PDAs could directly receive and evaluate the sensor values, and could request the needed RFID data. However, in the setting described, there is no communication channel shared by RFID readers and PDAs. Although both have WLAN access, the RFID readers use encrypted communication whereas the PDAs do not. Consequently, their alternative composition plan would be falsified by the verification tool and rejected in the implementation process.

Ontologies for Service-to-Device Mapping

In addition to manual deployment using dedicated deployment tools, the smart items middleware should be able to automatically map required services onto the most suitable available devices, or to relocate services from one device to another based on changing system conditions. Such automatic deployment of services is essential to manage the increasing complexity of distributed large-scale smart items solutions or scenarios, where available devices change frequently.

An automatic “intelligent” mapping of required services to the most suitable available device requires a formal description of the service requirements (e.g., CPU power, available main memory, or communication means) as well as the capability profiles of the available target devices. Ontologies can provide the formal description means for the specification of service requirements (which can be stored together with the semantic service description in the service repository) and the device profiles (cf. benefit unambiguous, formal representation of shared meanings of terms). Based on this foundation, reasoning can be used to
specify appropriate metrics to identify the “best” target device for a given service to be deployed (cf. benefit reasoning saves development efforts). Note, however, that significant modeling efforts (for the ontologies as well as for the rules) have to be expended before the benefits can be reaped (cf. limitation development vs. modeling efforts).

In our running example, services of the verified composition plan are mapped to appropriate devices. Services chosen for the gateways, RFID readers, and PDAs are to be installed on all those devices that have appropriate locations and properties (e.g., sufficient memory). Also, at this point, deployment constraints are specified (e.g., that the temperature filter service gets deployed to only 30% of the nodes, which are also required to have a certain amount of battery power left). This constraint imposes the need for a remapping of services if some sensor nodes are running out of power. The service mapper then maps the given services to the most suitable devices available by taking into consideration a formal specification of these mapping constraints and the service and device profiles. The actual service deployment is done by a service injector component which installs the respective service binaries for the respective platform.

**CONCLUSION**

The chapter introduced the reader to ontologies with the focus on the recent W3C recommendations RDF(S) and OWL that, for the first time, provide standard ontology languages. The recommendations have high impact on future tools and the interoperability of ontology-based applications. We further contrasted the pros and cons of ontologies on a general level and showed where ontologies can be used to counter fundamental challenges to enable scalable services-based ubiquitous computing applications in real-time enterprises.

**FUTURE RESEARCH DIRECTIONS**

One future research direction of this work is the collaboration between the nodes of the sensor network. By using collaborative algorithms, the nodes are able to perform business processes autonomously and independently from the back-end, only reporting exceptional situations (such as a dangerous constellation or the change of a monitored value). The expected outcome is a system that is highly autonomous, fault-tolerant, and scalable. Powerful tools will be developed, that allow business professionals to model the business logic to be executed by the nodes, simply by combining basic services.

Another area of future work is concerned with closing the information loop for the manufacturer of a product over the whole product lifecycle. At the beginning of the life of a product (including design and production), the manufacturer has detailed information available. However, once the product is delivered and used (middle of life), as well as at the end of its life (when it is recycled or refurbished), the information available to the manufacturer is scarce. By attaching product embedded information devices (PEIDs, such as rewritable RFID-Tags, embedded systems, or sensors) to the goods that can gather and store information during their lifetime, the information loop can be closed. This will allow business applications that are integrated in complex business processes to uniformly access data stored on or gathered by PEIDs in real time. This also enables the applications to configure and control PEIDs and update their memory.

Finally, there is a body of work in research and academia around what is called Semantic Web Services (McIlraith, Son, & Zeng, 2001). The principal objective of Semantic Web Services is a wide-reaching formalization that allows full automation of the Web service management tasks, such as discovery and composition. The core of the proposals lies in creating semantic standards
for the markup of Web services. Hence, another research direction is to evaluate whether the existing proposals can be used to also counter the technical challenges of our domain. The existing proposals are OWL-S (Martin, Burstein, Hobbs, Lassila, McDermott, & Mellraith, 2004), WSMO (Fensel & Bussler, 2002), and WSDL-S (Akkiraju, Farrell, Miller, Nagarajan, Schmidt, & Sheth, 2005).

REFERENCES


ADDITIONAL READING


Ontologies for Scalable Services-Based Ubiquitous Computing


ENDNOTES

1 Depending on the specific language used, classes are also called concepts or universals, relations are also called properties or roles, and instances are also called individuals or particulars.

2 Note that many of these primitives are also part of the OWL Lite language. Furthermore, Web compliance as described above holds for all three OWL variants.

This work was previously published in the Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises, edited by M. Mühlhäuser and I. Gurevych, pp. 88-106, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 6.12
Web Personalization for E-Marketing Intelligence

Penelope Markellou
University of Patras, Greece, & Research Academic Computer Technology Institute, Greece

Maria Rigou
University of Patras, Greece, & Research Academic Computer Technology Institute, Greece

Spiros Sirmakessis
Technological Educational Institution of Messolongi, Greece, & Research Academic Computer Technology Institute, Greece

ABSTRACT

As the Web is growing exponentially, online marketing has been changed by the newly provided technological capacities and digital channels of sales. Online marketing or e-marketing is the adaptation and development of marketing strategies in the Web environment and includes all factors that affect a Web site's efficiency, like the idea, the content, the structure, the interface, the implementation, the maintenance, the promotion and the advertising. Since more and more businesses are using the Web to conduct their activities, issues like interface usability, easy navigation and effective supporting services become critical and influence their success dramatically. However, one important problem that arises is that Web users are confronted with too many options. Currently, Web personalization is the most promising approach to alleviate this information overload and to provide users with tailored experiences. It improves user interaction with Web sites and offers them the ability to establish long-term and loyal relationships. The scope of this chapter is to give a comprehensive overview of research issues on personalized e-marketing applications. We focus on the importance of personalization as a remedy for the negative effects of the traditional “one-size-fits-all” approach. Next, we explore the different steps of the personalization process providing information about interesting research initiatives and representative commercial tools for producing personalized Web experiences. Finally, we demonstrate the close relation between per-
sonalization and Web mining and discuss open research issues.

INTRODUCTION

In today’s competitive business world, understanding and leveraging the power of the four Ps of marketing, namely product, price, place and promotion has become one of the major factors that determines a company’s marketing future survival or success. However, these four Ps have all been modified by technological evolutions and especially the Web and they have been converted into the four Ps of e-marketing, namely permission, privacy, profiling and personalization (Goldsmith, 1999). According to (Mobasher & Dai, 2003) “personalization is the provision to each individual of tailored information, products, or services” and the subject of this chapter is to study the significant role of personalization in intelligent e-marketing applications.

The roots of personalization are traced back to the introduction of adaptive hypermedia applications in Brusilovsky (1996, 2001). Adaptive hypermedia were introduced as an alternative to the traditional “one-size-fits-all” approach, building a model of the goals, preferences and knowledge of each individual user, and using this model throughout the interaction, in order to adapt to the user’s specific needs (Perkowitz & Etzioni, 1997, 2000a).

Personalization is a broad scientific and technological area, also covering recommender systems, customization, one-to-one marketing, and adaptive Web sites (Blom, 2000; Mulvenna et al., 2000; Schafer et al., 2001). It is the process of gathering and storing information about Web site visitors, analyzing the information in order to learn users’ patterns, habits and preferences, and based on this analysis, delivering the right information to each visitor at the right time (Eirinaki & Vazirgiannis, 2003).

Personalization can be deployed by e-businesses (any Web site whose operation is related to commercial purposes, e.g., e-shop, e-procurement, e-auction, e-mail, portal, etc.) and be used as the key technology for managing customer relationships, running targeted advertisement campaigns, promoting products or services and customizing Web site content (Perner & Fiss, 2002). Besides, according to the forth of the five mutable laws for Web marketing (Wilson, 1999) the objective of e-commerce sites should be to “pull people to the site by its attractive content, then push quality information to them regularly via e-mail.” However, personalized content apart from e-mail can be advertising, recommended items, screen layout, news, or anything else accessed via commercial Web site.

The digital channels that can be deployed by an e-business to reach its customers are numerous: Web, e-mail, ftp, chat, search, voice over IP, interactive multimedia, etc. The power of these digital channels is significant, since they can dramatically change the way the e-business listens, understands and responds to its customers and suppliers. Moreover, these channels facilitate instantaneous responses to the information gathered by various touch-points (buttons’ selection, hyperlinks, transactions, log files, etc.) without the need for human intervention (Yao et al., 2001).

Web personalization may be implemented in the form of:

- personalized content presentation and delivery (inserting or removing thematic units/sections/paragraphs, optional explanations or detailed information, personalized recommendations/offers/prices/products/services).
- personalized structure (sorting, hiding, unhiding, adding, removing or highlighting links), e.g., yahoo.com (Manber et al., 2000).
Web Personalization for E-Marketing Intelligence

- personalized Web site layout presentation and media format (from images to text, from text to audio).

The above can be based on the analysis of data from online and off-line sources using a variety of techniques: association rules mining, clustering algorithms, classification techniques, collaborative filtering, patterns discovery, prediction, log files analysis, etc. (Vassiliou et al., 2002).

The benefits for the businesses can be significant and include among others increasing site usability, converting users to buyers, retaining current customers, re-engaging customers, penetrating new markets, etc. (Nielsen, 1994). On the other hand, personalization applied to the e-commerce domain is an effective way to provide added value to customers by knowing and serving them as individuals. Customers need to feel that they have a unique personal relationship with the businesses they come in contact with. Users have now the ability to visit Web sites that allow them to find information or choose products and services to buy fast and easily. Moreover, they can receive e-mail, newsletters or other information that they attach importance to. And even though a significant percentage of online users give their permission to receive these kinds of information from marketers, they tend to unsubscribe and take back their permission when their preferences and needs are not met and they end up receiving useless or irritating information. This observation brings up the importance of quality and precision in producing and delivering personalization, as a means for assuring customer faith and loyalty.

There are two general approaches to personalization in e-commerce applications:

- **Buyer driven**: the customer subscribes to different services, fills in forms or questionnaires, rates products, participates in surveys, etc.
- **Seller driven**: the adaptation is controlled by the owner of the e-shop. Content is served up using pre-defined business rules, targeted online advertising, as well as product cross-selling and up-selling. It typically uses a rules engine which captures and encodes business rules, describes workflows and automates manual tasks.

The basic steps comprising the personalization process (depicted in Figure 1) are the following (Cooley et al., 1999a; Srivastava et al., 2000):

- Data collection.
- Data processing.
- Personalization output.

![Figure 1. Personalization process](image-url)
This chapter aims to define Web personalization for intelligent e-marketing applications, describe the tasks that typically comprise the personalization process and demonstrate the close relation between personalization and Web mining. Moreover, our goal is to illustrate the future trends in the field and, in this way, suggest directions that may trigger new scientific research.

**DATA COLLECTION**

By understanding consumers’ behavior, more efficient e-marketing strategies will become available to drive Internet use and e-commerce applications. Marketing efforts intended to enhance Web site use are expected to follow a two-fold strategy:

- Turn non-users into users.
- Expand usage of current users.

Consumers supported by a personalized system will be more likely to either turn from non-shoppers into active shoppers or increase their previous shopping volume. Web sites need to encourage users to discuss problems, and use this feedback to improve both products and services. Web sites should try to collect customer information and use that information to develop a relationship with customers. Customer satisfaction is the key for customer retention. Like traditional stores, online stores also need to build strong relationships with their customers. Technology provides many advantages over traditional ways of business with commercial Web sites using techniques such as online user groups, input from previous customers (rankings, comments, opinions, product assessments, etc.), order tracking and more.

So, the first step in the personalization process is the acquisition of data about the users (a task that is in continuous execution in most of the cases). User data must be transformed into an internal representation (modeling) that will allow for further processing and easier update. Indeed, personalization in order to produce satisfactory results needs different kinds of data. Some data can be observed by the system while other have to be provided by the user. The collection of information that describes a particular user is called a user profile, and consequently, a good such model comprises the basis for personalization activities. These profiles may be static or dynamic based on whether -and how often- they are updated. More specifically, the information incorporated in a use model may include: the user’s identification profile, the preference profile, the socio-economic profile, user’s ratings, reviews and opinions, the transaction profile, the interaction profile, the history profile, etc.

There are two general methodologies for acquiring user data depending on whether the user is required to be actively engaged in the process or not:

- **Reactive approach**: the user is asked explicitly to provide the data using questionnaire forms, fill-in preference dialogs, or even via machine readable data-carriers, such as smart cards.
- **Non-reactive approach**: the system implicitly derives such information without initiating any interaction with the user using acquisition rules, plan recognition, and stereotype reasoning.

Static profiles are usually acquired explicitly while dynamic ones are acquired implicitly by recording and analyzing user navigational behavior. In both approaches, we have to deal with different but equally serious problems. In the case of explicit profiling, users are often negative about filling in questionnaires and revealing personal information online, they comply only when required and even then the data submitted may be false. On the other hand, in implicit profiling, even though our source of information is not biased by the users’ negative attitude, the problems encountered derive once again from the
invaded privacy concern and the loss of anonymity. Personalization is striving to identify users, record their online behavior in as much detail as possible and extract needs and preferences in a way they do not notice, understand or control. The problem of loss of control is observed in situations where the user is not in control, a problem known as loss of control (Kramer et al., 2000; Mesquita et al., 2002; Nielsen, 1998).

Moreover, to maximize data gathering opportunities the Web site should collect data from every customer touch point, online and off-line. Online customer touch points include:

- **Registration:** the Web site asks some basic information about the customer (e.g., name, address, phone number, fax, interests, preferences, etc.), including the e-mail address and the password. Being a registered user makes future purchases faster, easier and friendlier.
- **Transactions:** purchase data or information requests.
- **Sign-ups:** newsletters, e-mail notifications, samples, coupons, partner offers, etc.
- **Customer profiles or user preferences.**
- **Customer surveys:** research-related and entertaining content surveys.
- **Customer service.**
- **Web log files:** pages viewed, categories searched, links clicked, etc.
- **Incoming and outgoing URLs (URL linking to the store, and links leading outside the store).**
- **Advertising banners.**
- **Sweepstakes** and other promotions requiring customer data.

Off-line customer touch points on the other hand may comprise:

- Customer service by phone, stored in the customer profile database.
- In-store transactions (meaning physical store purchases).
- Various surveys.
- Paper submissions (e.g., sweepstake or promotion entries).

Perhaps, the most important data source is the initial registration. In most cases this registration process is more important than the first transaction, in that the act of registering indicates that a customer wants to start a “conversation” or a relationship and gives the store permission to begin this process. When adequate data is collected, subsequent interactions with the store may well exceed the visitor’s expectations. Ensuring that the store allows customers to update and modify their own profile data not only will keep the customer information up-to-date, but it will also engender more trust because customers know what information is maintained about them by the e-store.

Another equally effective way to gather data about the customer is when the system does not explicitly ask for any information at all. Many successful Web sites use cookies and unique identifiers to make customer-specific data collection invisible to the customer.

Different kinds of data are used in personalization process:

- data about the user.
- data about the Web site usage.
- data about the software and hardware available on the user’s side.

**User Data**

This category denotes information about personal characteristics of the user. Several such types of data have been used in personalization applications. One source of information affecting customers’ decision-making process and attitudes is their **demographic traits**. These traits include...
name, address, zip code, phone number, other geographic information, gender, age, marital status, education, income, etc. All customers are not equal. Different customers and customer segments value different things, so for some it is important for a Web site to provide lower prices and faster delivery, while for others the priority focuses on quality, number of choices and convenience. An example found in Liebermann and Stashevsky (2002) reveals differences in attitudes based on sex and according to it males worry more than females for the vast volumes of Internet advertising.

Another source of information relates to user’s knowledge of concepts and relationships between concepts in the application-specific domain (input that has been of extensive use in natural language processing systems) or domain-specific expertise.

Moreover, valuable types of data may be user skills and capabilities in the sense that apart from “what” the user knows, in many cases it is of equal importance to know what the user knows “how” to do, or even further, distinguish between what the user is familiar with and what he/she can actually accomplish.

Finally, interests and preferences, goals and plans are used by plan recognition techniques where identified goals allow the Web site to predict interests and needs and adjust its contents’ structure and presentation for easier and faster goal achievement.

Usage Data

Usage data may be directly observed and recorded, or acquired by analyzing observable data (whose amount and detail varies depending on the technologies used during Web site implementation, i.e., Java applets, etc.), a process known as Web usage mining (Markellou et al., 2004, see also section “Personalization and Web Mining”). Usage data may either be:

- Observable data comprising selective actions like clicking on an link, data regarding the temporal viewing behavior of users, ratings (using a binary or a limited, discrete scale) and other confirmatory or disconfirmatory actions (making purchases, e-mailing/saving/printing a document, bookmarking a Web page and more), or
- Data that derive from observed data by further processing (measurements of frequency of selecting an option/link/service, production of suggestions/recommendations based on situation-action correlations, or variations of this approach, for instance recording action sequences).

Environment Data

On the client side, the range of different hardware and software used is large and keeps growing with the widespread use of mobile phones and personal digital assistants (PDAs) for accessing the Web. Thus in many cases the adaptations to be produced should also take into account such information. Environment data addresses information about the available software and hardware at the client side (browser version and platform, availability of plug-ins, firewalls preventing applets from executing, available bandwidth, processing speed, display and input devices, etc.), as well as locale (geographical information that can be used, for instance, to automatically adjust the language, or other locale specific content, such as the local time or the shipping costs).

DATA PROCESSING

The step following after the collection of data from all sources and the building of users’ profiles is data processing. Initially, some preparation activities take place in order to clean the data and facilitate their manipulation. For instance, entries that do not reveal actual usage information are
removed and missing data are completed. Then follows the application of statistical and data-mining techniques in order to detect interesting patterns in the pre-processed data. The most well-known techniques that are used for data analysis include clustering, classification, association rules mining, sequential pattern discovery and prediction. A more detailed description of each technique follows.

**Clustering**

*Clustering algorithms* are used mainly for segmentation purposes. Their aim is to detect “natural” groups in data collections (e.g., customer profiles, product databases, transaction databases, etc.). They compute a measure of *similarity* in the collection in order to group together items that have similar characteristics. The items may either be users that demonstrate similar online behavior or pages that are similarly utilized by users. The produced groups (database segmentation into clusters of similar people, e.g., customers, prospects, most valuable or profitable customers, most active customers, lapsed customers, etc.) can be based on many different customer attributes (e.g., navigation behavior, buying behavior or demographics). There are several clustering algorithms available: Hierarchical Agglomerative Clustering or HAC (Rasmussen, 1992; Willett, 1988), k-means clustering (MacQueen, 1967), Self-Organizing Maps (SOMs) or Kohonen (1997).

**Classification**

The main objective of *classification algorithms* is to assign items to a set of predefined classes. These classes usually represent different user profiles and classification is performed using selected features with high discriminative ability as refers to the set of classes describing each profile. For example the profile of an active buyer can be:

- sex = male
- 30 ≤ age ≤ 40
- marital-status = single
- number-of-children = 0
- education = higher

*Figure 2. Clustering*

---

**Initial data collection**

Data clusters based on similarity attributes
Web Personalization for E-Marketing Intelligence

Figure 3. Classification

Association Rules

Association rules connect one or more events. The aim is to find out associations and correlations between different types of information without obvious semantic dependence. In the Web personalization domain, this method may indicate correlations between pages not directly connected and reveal previously unknown associations between groups of users with specific interests (Agrawal et al., 1993; Agrawal & Srikant, 1994; Chen et al., 1996, 1998). Such information may prove valuable for e-commerce and e-business Web sites since it can be used to improve Customer Relationship Management (CRM). Some examples of association rules are the following:

- 20% of the users that buy the book “Windows 2000” also select “Word 2000” next,
- 50% of the users who visited the “Help” pages belong to the 25-30 age group,
- 30% of the users who accessed the Web page “Special Offers” placed an online order for product “DVD – Lord of the Rings,”
• 60% of the users who ordered the book “Harry Potter” were in the 18-25 age group and lived in Athens,
• or 80% of users who accessed the Web site started from page “Products.”

Sequential Pattern Discovery

Sequential pattern discovery is an extension to the association rule mining technique, and it is used for revealing patterns of co-occurrence, thus incorporating the notion of time sequence. A pattern in this case may be a Web page or a set of pages accessed immediately after another set of pages. Examples of sequential patterns can be:

• “45% of new customers who order a mobile phone will spend more than 50 Euros using it within 30 days.”
• “Given the transactions of a customer who has not bought any products during the last three months, find all customers with a similar behavior.”

Prediction

Predictive modeling algorithms calculate scores for each customer. A score is a number that expresses the likelihood of the customer behaving in a certain way in the future. For example it answers questions such as:

• What is the possibility of a user to click on a certain banner?
• How likely is a user to re-visit the Web site in the next month?
• How many orders will be placed by customers from abroad?

These scores are calculated based on rules that are derived from examples of past customer behavior. Predictive Modeling methods include
Decision Trees, Regression Models and Neural Networks.

The prediction can be:

- **Off-line**: the decisions are pre-calculated and picked-up at the start of a repeated visit (the click-stream of the current visit is not used).
- **Online**: the current visit’s click-stream is used for decisions.

Moreover, predictive models can be used in order to decide the content to be displayed to a user (which constitutes a crucial part of CRM).

**PERSONALIZED OUTPUT**

After data processing, the extracted knowledge has to be converted into intelligent information, interaction, or interface for each customer. Every segment or cluster is a separate e-marketing opportunity that should be fulfilled differently. This section focuses on the output of e-marketing personalization: content, structure and presentation adaptations based on the constructed user, usage and environment models, as well as the resulting secondary inferences.

**Personalized Content**

This type of adaptations is applied to the informational content of a Web site. The information is filtered and then presented to the users. This can have many forms including additional or reduced information, optional explanations, recommendations, offers, special prices, products, services, news, e-mails, help, etc. The techniques used for producing such adaptations include inserting, removing, altering, sorting, coloring, dimming text fragments, using stretch-text, etc.

**Personalized Structure**

This type of adaptations relates to changes in the link structure of a Web site. It is important for the user to see only the information of interest. Here, we refer to changes that can be applied to the link structure of Web pages, sections, categories, etc., as well as to recommendations for products, services, information, navigation, etc. The techniques used for producing such adaptations are sorting,
hiding, unhiding, adding, removing, enabling, disabling, or highlighting links.

**Personalized Presentation**

This type of adaptations refers to changes that affect not the content itself, but its format and layout (e.g., from images to text, from text to audio, from video to still images) and it is mainly used for Web access through PDAs or mobile phones, or in Web sites that cater to handicapped persons.

**Recommendations**

Although recommendations belong to the personalized content output category, they are examined separately because they feature as an important component of e-commerce Web sites. Their purpose is to provide access to specific items through promotional links, such as those that are supplied by cross-selling or up-selling options.

- **Cross-selling**: it suggests products related to the one(s) the user is currently viewing. In many cases, these are complementary products. For example, proposing a music CD with a book or batteries with toys.
- **Up-selling**: it suggests products perhaps more expensive or advanced to the one(s) the user has chosen. The customer will be informed about products available in the next (upper) price level, which he/she may not have known about. The degree of applicability of this tactic depends on the type of products, and this also applies to cross-selling and up-selling as well.
- **Opinions of other customers**: it suggests additional products that the customer may also like to purchase, based on what other customers (considered as like-minded) have bought.
- **History data**: analyzing the history of past purchases (stored in a transaction database), the e-shop is able to offer customers a targeted range of choices that are most likely to fit their profile.

It is worth mentioning that the home page of a Web site is a good place to put items in order to be promoted, as it is a place that draws by default user attention. This page should be constantly updated to keep people coming back. A good Web site design should provide an option (in the administrative page) that allows the store manager to determine whether a product is to be treated as a featured product appearing on the home page, or not.

**PERSONALIZATION AND WEB-MINING**

A big part of Web personalization as we experience it nowadays is based upon the core technologies of Web mining. In fact, in this chapter we have referenced a number of Web mining techniques (clustering, classification, association rules mining, sequential pattern discovery and prediction) when discussing the phase of data processing. Web mining is broadly defined as “the use of data mining techniques for discovering and extracting information from Web documents and services and it is distinguished as Web content, structure or usage mining depending on which part of the Web is mined” (Kosala & Blockeel, 2000). It is a converging research area using techniques and methods that derive from various research fields such as: databases (DBs), information retrieval (IR), artificial intelligence (AI), as well as psychology and statistics.

The distinctions between the three main categories of Web-mining are not clear-cut. Web content mining might utilize text and links and even the profiles that are either inferred or explicitly inputted by users. The same is true for Web structure mining—it may use information about the links in addition to the link structures. Or, the traversed links can be inferred from the pages that
were requested during user sessions and can be found recorded in server logs (which is a typical Web usage mining task). In practice, the three Web-mining categories can be used in isolation or combined in an application, especially in Web content and structure mining since links may be considered as part of the content of a Web document (Chakrabarti et al., 1999). In the majority of cases, Web applications base personalization on Web usage mining, which undertakes the task of gathering and extracting all data required for constructing and maintaining user profiles according to the “logged” behavior of each user.

Web industry and researchers from diverse scientific areas have focused on various aspects of the topic. There are many research approaches and commercial tools that deliver personalized Web experiences based on business rules, Web site content and structure, as well as the user behavior monitoring. The most well-known applications of Web personalization at a research level include: Letizia (Lieberman, 1995), WebWatcher (Armstrong et al., 1995; Joachims et al., 1997), Fab (Balabanovic & Shoham, 1997), Humos/Wifs (Ambrosini et al., 1997), SiteHelper (Ngu & Wu, 1997), Personal WebWatcher (Mladenic, 1999), Let’s Browse (Lieberman et al., 1999), SpeedTracer (Wu et al., 1998), WebPersonalizer (Mobasher et al., 2000), WebSIFT (Cooley et al., 1997, 1999b; Cooley et al., 2000), Web Utilization Miner - WUM (Spiliopoulou & Faulstich, 1998; Spiliopoulou et al., 1999a, 1999b; Spiliopoulou & Pohle, 2000), MIDAS (Buchner et al., 1999), IndexFinder (Perkowitz & Etzioni, 2000b).

Moreover, many vendors provide a variety of commercial tools that support mining for Web personalization. These tools can be integrated directly into a Web site server in order to provide users with personalized experiences.


CONCLUSIONS

The explosive evolution of the information technology and the Internet has introduced enormous changes in the way marketing strategies are being
implemented. This growth was followed by problems concerning effective and efficient methods and techniques for supporting consumers and providing them with tailored Web experiences.

Personalization has the potential to be the answer and e-marketing should further focus on it. This technique can be applied on different dimensions of a general Web site marketing strategy, including: sales, advertisement, communication, customer support, etc. In this way any e-business, i.e., information portal, e-commerce site, or e-learning system can improve its performance by addressing the individual needs and preferences of each user, increasing satisfaction, promoting loyalty, and establishing one-to-one relationships.

Many research approaches, initiatives and techniques, as well as commercial tools provide Web personalization based on business rules, Web site contents and structuring, user behavior and navigational history as recorded in Web server logs.

Customers expect to be able to interact with an e-commerce site, find information and make online purchases easily and effectively. On the other hand, e-businesses expect to sell as many products and services as possible easily and cost-effectively as well. Thus, an effective e-marketing strategy should be considered from both the perspective of the customer and the e-business owner.

Specifically, from the customer’s point of view, the Web site should ensure that:

- Efficient recommendations about new products, product discounts, special offers, coupons, sweepstakes, etc., are produced and delivered.
- All second-time visitors receive personalized content to meet their needs and that this content is embedded into recommendations about additional relevant products to be purchased.
- Browsing through the products of the online catalogue is easy.
- Useful assistance is available during all steps of the shopping process.
- Customers’ time is respected and irritation is minimized.
- Customers’ individuality is respected.
- Customers are allowed to specify and modify their own profile settings.

As regards the standpoint of the e-business, the Web site should:

- Promote the products effectively.
- Maintain customer profiles. This means that the system should capture customer behavioral information from the login procedure, completed purchases, response to questionnaires, etc. The data should be updated (if not dynamically) at predefined time intervals.
- Generate consumer demographics.
- Manage navigation data.
- Analyze sales data.
- Gather statistical data.
- “Remember” and analyze everything the store needs to “know” about the customer.
- Encourage customers to “leave something of themselves behind” by multiple and “invisible” methods.
- Increase second-time purchases (in the sense that the customers should keep coming back and thus develop a sense of loyalty towards the e-store).

An open issue in the field relates to the fact that in order to fulfill its aims, personalization needs rich data about users (e.g., demographics, behaviors, structure, etc.). If the data are poor, the produced personalized output will be unsuccessful in understanding and satisfying user needs and goals.

In most cases, users are negative about revealing personal information and resist to the idea of being stereotyped. Therefore, another issue that must be taken into consideration when planning
and implementing personalization is privacy (Earp & Baumer, 2003; Kobsa & Schreck, 2003; Markellos et al., 2004). Tracking users’ behavior means collecting (potentially sensitive) information about them. Many research and commercial tools are available ensuring to some degree that privacy is not jeopardized (Cingil et al., 2000). However, if customers resist doing so, they will not be able to join the personalized segment.

Finally, a relatively recent development that is foreseen to greatly affect Web personalization is the expansion of the semantic Web. Semantic Web mining combines the two fast-developing research areas of semantic Web and Web mining with the purpose of improving Web mining by exploiting the new semantic structures in the Web. Berendt et al. (2002) gives an overview of where the two areas meet today, and sketches ways of how a closer integration could be profitable.

Summarizing, in this chapter we presented the importance of personalization for intelligent e-marketing Web site applications. We explored the different phases of the personalization process (data collection, data processing and personalization output), we demonstrated the close relation between personalization and Web mining and concluded by discussing some open issues in the field.

REFERENCES


Web Personalization for E-Marketing Intelligence

shop on Machine Learning in User Modelling of the ACAI99, Chania, Greece.


*This work was previously published in Contemporary Research in E-Marketing, Volume 1, edited by S. Krishnamurthy, pp. 48-68, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
INTRODUCTION

Knowledge sharing in today’s distributed organizations is a challenge. Hierarchical structures may not support the fast flow of information or the efficient co-creation of knowledge from specialized and tacit individual knowledge bases (see Grant, 1996; Miles et al., 2000; Adler, 2001). There is therefore a need to devise new patterns for leveraging dispersed knowledge within organizations and across organizational borders. In the following we explore internal company weblogs as a potential new channel for sharing knowledge and expertise.

Knowledge is deeply embedded in social interaction. Recent advances in ICT have led to the production of social software, the primary purpose of which is knowledge sharing (Teigland & Wasko, 2005). One example of this is the emergence of weblogs (or blogs), personal Web pages that incorporate regular posts in reverse chronological order about a particular topic, current events or personal thoughts and expression (Blanchard, 2004; Herring et al., 2004; Wagner & Bolloju, 2005). Our focus is on internal knowledge blogs. A small proportion of blogs could be classified as knowledge blogs, implying the online equivalent of professional journals in which authors share new knowledge in their professional domains: they report on their research progress, share references and make observations (Wagner & Bolloju, 2005). However, the application of blogs in knowledge management is a new phenomenon, and firms have been slow to do so. Current research fails to promote understanding of the applicability of blogs in networked teams and organizations, and
of their role in the social processes of knowledge sharing and creation.

Thus we contribute to the emerging literature by exploring the critical factors involved in applying internal company blogs. Most studies have addressed only the rational part of mediated communication, that is, the characteristics of communication technologies and their capacity for managing and transferring information. Our aim in this article is to include the social perspective, and to provide a categorization that combines the rational-social and individual-organizational dimensions. We then report on a mini-case study in order to illustrate the identified factors. We argue that anyone wishing to exploit the potential of knowledge sharing through internal blogs should understand their nature and characteristics in order to be able to make the best possible media choices.

BACKGROUND

Knowledge Blogs

Knowledge blogs are used in internal communication for knowledge-sharing purposes (Zerfass, 2005): to promote a shared understanding of the perspectives of CEOs and employees (Dearstyne, 2005), to coordinate projects and present ideas, and to bridge the gaps between individuals with various views in order to identify new trends in innovation communities (Wagner & Bolloju, 2005). Knowledge sharing, collaboration and best practices are endemic in blogs (Ojala, 2004).

Knowledge blogs appear at the intersection of professional communication and personal reflection. They serve as an ideal medium for experts sharing an interest in a specific topic: they enable storytelling, reflection, the capturing of information, and the sharing thoughts or ideas, thus making individuals’ “hidden” processes more visible and promoting both personal and collective learning (Röll, 2004; 2006). Thus, they can offer business representatives a means for expanding the boundaries of knowledge sharing and creation (Herring et al., 2004). Blogs as a social medium facilitate the sharing of both objectified (codified) and collective knowledge (social/organizational practices and experiences).

Approaches to Computer-Mediated Communication

Theories of computer-mediated communication (CMC) have been categorized as rational or social (see Kock, 2005; Webster & Trevino, 1995). The former rely on the hypothesis of rationality and effectiveness, involving users who select the communication medium of appropriate richness and level of social presence for each task. According to the early theories regarding the effects of CMC, the “richer” the media are (i.e., the more they allow nonverbal cues, provide immediate feedback, and convey personality traits), the better suited they are to human-to-human interaction (Short et al., 1976; Daft & Lengel, 1986).

From a social perspective, alternative approaches concentrating on the applicability of CMC have emerged, including the social-influence model devised by Fulk et al. (1990), social information processing theory as developed by Walther (1996), and Lea and Spears’ (1992) theory of social identity and de-individuation (SIDE). According to the social-influence model, the technological features are not inherently decisive in the choice of communication media, but they are influenced by social-group norms and membership (Fulk et al., 1990; Markus, 1994). Walther (1996), in turn, suggests in his theory of social information processing that the “cues-filtered-out” conditions in CMC do not prevent relational communication, although social relationships take a longer time to develop in computer-mediated groups than in face-to-face groups. The SIDE theory concerns processes of social identification and self-categorization, that is, in-group and out-group membership (Lea & Spears, 1992). Attention should be paid to
these processes given that similarity with group members is believed to strengthen the individual's identity, which in turn has a positive effect on their willingness to co-operate (Kramer et al., 1996; Järvenpää & Leidner, 1999).

THE MAIN FOCUS OF THE ARTICLE

In the following we discuss the critical factors involved in applying internal knowledge blogs. Some of the issues are naturally also related to other organizational communication media, but here we assess them in the context of blog communication, its defining characteristics and patterns of use in organizations.

The Nature of Blog Communication

We strongly believe that there should be conscious media choice in both firm-internal and firm-external knowledge sharing and creation (Robert et al., 2005). In order to exploit their concrete benefits, organizations should be aware of the strengths of blogs: they offer continuity, they concentrate on one or a few topics, and they promote efficient information management, the forming of networks around specific issues, the maintenance of shared narratives and the telling of stories (Cayzer, 2004; Röll, 2004; Wagner & Bolloju, 2005). They also facilitate the building of personal identity/status and simultaneously provide users with high connectivity, thus cultivating the development of online networks. As a medium, they are time and space flexible as they do not tie the receiver to a certain place or time. They also allow the receiver to process the message until it is wholly understood, and to search for extra information to ease understanding or to evaluate content and credibility. Blogs leave a record of the issues discussed, which is an advantage in business situations, and typically offer easy access to additional resources through hyperlinks, trackbacks (reverse hyperlinks) and other recommended blogs (Cayzer, 2004; Herring et al., 2004).

High-presence media seem to set demands on us that cannot always be fulfilled as we are not able to process so much information at once. Hence, Robert et al. (2005) came to the conclusion that rich media with a high social presence may ease the exchange of simple ideas, and hinder the exchange of more complex ideas (see Markus, 1994). This would imply that blogs could be well suited to more complex information exchange, to generating ideas, developing insights, and maintaining shared narratives (Wagner & Bolloju, 2005). On the other hand, they must attract the attention of the receiver: such low-presence media can be easily ignored (see Robert et al., 2005).

Individual Motivation and Competence

To be able to contribute through mediated communication, individuals have to possess the necessary skills and knowledge (Vartiainen et al., 2003). There is also the crucial question of motivation, as contribution requires time and effort (Kimble & Li, 2005). Articulating ideas through writing and storytelling seems to be one of the major motives for blogging (Nardi et al., 2004). It is also a means of self-regulated learning: reflecting on one's own learning and simultaneously accessing others’ reflections facilitate the development of a collective understanding of a topic (Baggetun & Wasson, 2006). In general, the main reason why people attach themselves to online social networks seems to be the motive to gain access to valuable information and knowledge for their personal benefit (Wasko & Faraj, 2000). It is probably employees’ intrinsic motivations that are decisive for active participation in blogging. However, employee motivation can also be enhanced by providing interesting content in the early stages of the implementation, and encouraging the free exchange of ideas (Nardi et al., 2004; Waterson, 2005).
The Role of Technology

In general, blogs are free, lightweight and relatively easy to apply, as blogging software offers a platform for Web publishing with no need for additional tools. They are also easy to maintain. Much of the knowledge is in text form, possibly enhanced by multimedia attachments (Herring et al., 2004; Kelleher & Miller, 2006). Syndication formats such as Really Simple Syndication (RSS) and Atom allow individual and community aggregators to collect, merge and sort blogging data from the ever-growing contents of the blogosphere (Cayzer, 2004).

Organizational Policies and Support

Kock (2005) stresses the fact that social influence may affect media choice, referring to the study conducted by Markus (1994) showing that employees used e-mail at the request of managers even if they perceived it as a “poor” communication medium. Support from managers and team leaders should be valued in the implementation of internal blogs. According to Holtz (2005), it is also important for policies on employee blogging to be clearly defined: the duality between personal reflections and professional status requires explicit norms. For instance, general guidelines covering each author’s personal responsibility should be provided—blogs convey individual thoughts and interactions, not corporate communication. Blogging policies should be made explicit before the implementation phase (Dearstyne, 2005) in order to prevent misuse of confidential or proprietary information, for example.

The Critical Mass of Users

The defining feature of communication technologies is that they require a certain proportion of users, or a critical mass, and when this has been achieved their use should spread rapidly: having a critical mass allows the users to reach the largest number of people with the least effort (Markus, 1987). Blog communication provides users with high connectivity, as it is easy to form social networks of people who share similar interests. From the rational perspective, efficient communication still requires a number of bloggers who regularly post their thoughts and comment on others’ posts. On the other hand, growth in volume becomes a critical issue, as blogging networks require a balance between enough connections and the amount of information: one individual can reasonably manage only a certain number of community ties.

Organizational Culture

In order for knowledge sharing to succeed and be useful, organizations must value open-minded and non-hierarchical idea exchange (Wagner & Bolloju, 2005), and share a common ground with membership, identity and communication procedures, as the organizational culture is the binding factor for collaboration (Huq, 2005; Kimble & Li, 2005).

Kelleher and Miller (2006) considered the potential advantages of organizational blogs over Web sites in three experimental groups; they found that blogs were perceived to have a more conversational tone, which appeared as an appropriate relational maintenance strategy for virtual contexts. They also seemed to enhance identity-building and information richness through storytelling, which transfers personal and even tacit knowledge (ibid., Röll, 2004). Both personal and collective identities may be built up through blogging networks. Shared identity is critical for coordination as it lowers the costs of communication, and establishes explicit and tacit rules of coordination and influence. According to Kogut and Zander (1996), a shared identity leads to social knowledge that supports coordination, communication and learning: it implies a moral order, rules of exclusion, and norms of procedural justice. Some behavioral studies show the
critical role of group *membership* in attributing the “insiders” and “outsiders”, and also suggest that attribution could ease potential dissonance between opportunism and loyalty (ibid., Kramer et al., 1996).

**Trust**

A lack of trust may prevent knowledge sharing (Kimble & Li, 2005). In general, trust is a multi-level concept. First, it takes self-confidence to be willing to express oneself openly and to reveal one’s identity through blogging. Trust in the information and in the competence of the author(s) is also required. Knowledge has to come from a trusted source if it is to have a desirable effect on the receiver: processes of word-of-mouth are inherent in blogging (see Torio, 2005). Sufficient trust in the enabling technology is also needed, regardless of whether it is open-source software or a standardized company offering. Finally, the employee needs to be confident enough to express her/himself through blogs despite the possible threats and sanctions, such as being punished for making local information transparent organization-wide (Dearstyne, 2005).

Figure 1 summarizes the ideas presented above in terms of the social-rational and individual-organizational dimensions in applying blogs. We propose that the rationally constructed factors provide the enabling structure for communication, while the actual outcomes are determined by both the social and rational factors.

**A Mini-Case of Applying Blogs in a MultiNational Firm**

The following is a real-life story of blogs in practice in a large ICT company. The mini-case demonstrates some success factors of knowledge blogs. We conducted a one-and-a-half-hour telephone interview with four representatives of the company in order to discuss why and how an internal blog platform had been implemented. The interview was tape-recorded and transcribed. Earlier, one of the interviewees had introduced his own wiki for research purposes. The use of wikis then expanded further, and in January 2005 an internal blog platform was taken into use. Initially, the organization did not support the development work: the evolution of wikis and blogs unfolded “bottom-up.” However, later the key developer was rewarded for his efforts.

The use of blogs was minimal at first, but it expanded due to easy implementation and use, word-of-mouth inside the company, and motivated people, who also blogged in their spare time. Thus the critical mass of users was eventually reached. “People realized the value of blogs …It serves as a means of internal marketing, i.e. ‘I know about this, so you can ask me’. This, in turn, reduces internal overlap.” Bloggers are able to trust each other’s ability, as blogs efficiently convey expertise. The existence of both intrinsic and extrinsic motivations (personal status, enjoyment, access to valuable knowledge for personal benefit, getting to know who knows what) was also implied: for the individuals concerned, blogging was a means of capturing knowledge, gaining status and forming relationships within the network. On the organizational level, blogs potentially increase efficiency.

The interviewees explicitly described the differences between various communication channels and awareness of the pros and cons of blogging. Due to their highly personal nature, blogs carry an informal and conversational tone. They offer a more “human voice” than other low-presence media. On the other hand, they represent “a shout in the dark, where the echo may or may not resound…”. In this respect, e-mail was also strongly contrasted to blogging: “E-mail is a channel of power, control and information-hiding. This conflicts with the blogging culture, which is the equivalent of open-source ideology. You want to share what you know.”

Thus the various cultures in organizations may conflict with each other and form barriers to the
use of knowledge blogs. “Very often people come and ask ‘Can I write this? Who can publish it?’ The Net is still seen as it was with us, a highly official system. People do not yet understand that blogging is about bringing the informal coffee-room discussions into another kind of social space.” Again, the coin has two sides: sometimes bloggers unwittingly broadcast confidential “coffee-room wisdom” and make it available organization-wide. It is difficult to define company blog policies, and they seem to form through trial and error.

The case could be summarized as follows. The process of internal knowledge blogs was initiated by a small core group, the members of which were aware of the opportunities and challenges of blogging, and were very familiar with technology-mediated social interaction. A critical mass of users was reached due to processes of word-of-mouth and related trust, employee motivation and the easy implementation of personal blogs. On the other hand, the blogging culture is in its early stages and there is still uncertainty about its role as a medium, and about personal authority in terms of making contributions. It is also difficult to identify highly confidential issues from “knowledge to be shared.”

DISCUSSION

Theoretically, our article contributes to the increasing understanding of how technology changes the way in which people interact socially and exchange knowledge with each other in a firm-internal context. We also categorized the critical factors involved in applying knowledge blogs for internal purposes. On a practical level, we presented a mini-case to illustrate the implementation process and the related critical factors.

The role of storytelling and expressing knowledge for its further construction and refinement could be essential in the transfer of organizational knowledge (Brown & Duguid, 1991; Lave & Wenger, 1991; Brown, 2001). Blogs seem to be a potential medium to be applied in firm-internal knowledge sharing and creation in networked and virtual organizations. They are conversational, inexpensive and easy to use, they allow empowerment and the free expression of thought, and they convey expertise and contribute to building personal and collective identities. However, identity-building may also lead to exclusive types of social networks if insufficient attention is paid to developing a trust-inclusive organizational culture, offering employee education on how to use blogs, and developing related organizational policies.

FUTURE TRENDS

In the emerging field of the knowledge-management applications of blogs, the success factors discussed above should be empirically validated and refined further. More research is also needed in order to enhance understanding of blogs as socio-technological innovations. Rice (1987) argues that, while computer-mediated communication systems process information about innovations, they are also an innovation that organizations must process, a phenomenon that has recently been illustrated in the development of blog communication. However, it is apparent from our emerging understanding of the co-evolution of innovations that organizations must also consider complementary and administrative innovations, such as organizational structures and processes that support the implementation of new communication technologies.

CONCLUSION

We identified several critical factors for applying internal knowledge blogs: understanding the nature of blog communication, individual com-
competence and motivation, the role of technology, organizational policies and support, a critical mass of users, the organizational culture, and trust.

Knowledge blogs may make organizational knowledge-sharing and related social networks more transparent through their ability to build identities, mediate both personal and collective knowledge, and serve as an institutional memory. They enhance storytelling-type of knowledge-sharing, and thus contribute to the transferring of tacit knowledge in networked and virtual organizations. They also have the potential to link the informal and the formal organization, which is a major challenge for organizational innovativeness (see Dougherty & Hardy, 1996). The formal organization may enhance the application of blogs by promoting open communication and a culture of sharing, while also providing technological and managerial support.

REFERENCES


Knowledge Blogs in Firm Internal Use


Knowledge Blogs in Firm Internal Use


KEY TERMS

Blog (or Weblog): A personal Web page including regular posts in reverse chronological order about a particular topic (Blanchard, 2004; Herring et al., 2004; Wagner & Bolloju, 2005).

Information Richness: Refers to the ability of the media to transfer cues and provide immediate feedback (Daft & Lengel, 1986).

Knowledge Blog: A weblog for experts sharing an interest on a specific topic and for documenting knowledge in their professional domains (Wagner & Bolloju, 2005).

Knowledge Management: Refers to a range of practices organizations use to identify, distribute and leverage knowledge to gain competitive advantage.

Media Choice: Refers to the examination and selection of communication channels, that is, in terms of media richness and the level of social presence (Robert et al., 2005).

Organizational Blog: A Weblog that is endorsed by an organization and maintained by its official or semi-official representative whose affiliations to the organization are public (Kelleher & Miller, 2006).

Social Presence: The degree and type of mediated interpersonal contact and intimacy (Short et al., 1976).
ENDNOTE

According to Markus, 1994, less rich media are often used by managers for handling complex tasks.

This work was previously published in the Encyclopedia of Networked and Virtual Organizations, edited by G.D. Putnik and M.M. Cunha, pp. 769-776, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 6.14
Towards Ambient Business:
Enabling Open Innovation in a World of Ubiquitous Computing

Christian Schmitt
University of Cologne, Germany

Detlef Schoder
University of Cologne, Germany

Kai Fischbach
University of Cologne, Germany

Steffen Muhle
University of Cologne, Germany

ABSTRACT

Ubiquitous computing enables the development of new innovative applications and services. Particularly influential on future business services will be the connection of the real with the virtual world by embedding computers or smallest processors, memory chips, and sensors into the environment and into physical objects, as well as using natural, multimodal customer interaction. Bearing in mind that ubiquitous computing entails the interaction of mobile smart objects and local smart environments hosted by different service providers, we propose an open approach to encourage the development of new and innovative smart applications and services. Considering the Open Source community, we assert that such an open approach reveals innovation potentials that cannot be achieved in proprietary information system environments. Most research projects, as well as commercial initiatives, however, focus mainly on specific, proprietary applications like smart shopping environments, and do not incorporate further prospects of using an open approach. Therefore, this chapter discusses as a first step the impact of Open Innovation in a world of ubiquitous computing from an economic perspective. Then, the design of an Open Object Information Infra-
Towards Ambient Business

structure (OOII) that enables Open Innovation in the context of ubiquitous computing is presented. Finally, an innovative smart service called the Federative Library (FedLib), which represents a first instantiation of the OOII, is introduced to demonstrate the feasibility of the design.

INTRODUCTION

The connection of the real and the virtual world encourages the development of new and innovative applications and services. Google is a prominent case which distinctly underlines the creative power of combining both worlds (Roush, 2005). In June of 2005, Google released an official API for Google Maps and Google Earth. Through this, consumers and external developers now have access to detailed aerial and satellite maps and advanced geographical visualization tools that can easily be used to display data atop the Google maps. Almost immediately, a community of “geotagging” map makers was formed that ties information on the Web to geographical coordinates in order to build geospatial applications (examples of so-called “map mash-ups” can be found in Gibson, 2006). Geotagging means amplifying physical places with information so that, for instance, users of mobile devices with location technologies can retrieve additional information related to their current locations. This includes location-based ads and listings for nearby shopping, dining, entertainment, and business outlets or even photographs, videos, stories, and other personal information uploaded to the Internet and pinned to specific latitudes and longitudes. As a result, the community participates in the innovation process and in building “a browser for the earth,” as John Hanke, general manager of Google’s Keyhole Group, describes Google Earth (Roush, 2005). This could enable a new and natural means of accessing location-based information. Wade Roush (2005) describes this vision by drawing a comparison: “Every page on the Web has a location, in the form of a URL. Now every location can have a Web page [...] it means that navigating both the Web and the real geography around us is about to become a much richer experience, rife with occasions for on-the-spot education and commerce. It means that we will be able to browse the Web-and the virtual earth encompassed within it-simply by walking around.” (p. 60)

As the Google case also illustrates, opening access to information systems and intellectual property (IP) to complementors, suppliers, users, and even to competitors encourages innovation, increases the attractiveness of platforms, and enriches service provision through the incorporation of external information sources. With the world evolving from Electronic Business to ubiquitous computing (Weiser, 1991) (and pervasive computing (Hansmann, 2003) or ambient intelligence (Aarts, 2001) respectively), even more comprehensive opportunities for implementing innovation are likely to arise. Innovative information gathering and access is enabled by the connection between the real with the virtual world by embedding computers or the smallest processors, memory chips, and sensors into the environment and into physical objects (Gershenfeld, 1999), as well as using natural, multimodal user interaction (Dam, 2001).

However, existing approaches frequently rely on rather proprietary information systems that reveal their data and services only partially or under certain conditions. Bearing in mind that ubiquitous computing entails the interaction of mobile smart objects and local smart environments hosted by different service providers, we propose using an open approach to encourage the development of new and innovative smart applications and services. Considering the Open Source community, we assert that such an open approach reveals innovation potentials that cannot be achieved in proprietary information system environments. Various developers with different perceptions can then search for relevant smart
services, as well as information gathered by smart objects or environments, combine them freely, and hence create their own innovative services.

The rest of the chapter is structured as follows. The following section presents related work. Then, we first introduce our framework for ambient business. We then discuss the impact and potential of Open Innovation in a world of ambient business from an economic perspective. Next, we present the design of an architecture called the open object information infrastructure (OOII) that enables open innovation in the context of ambient business. To demonstrate the feasibility of the design, we introduce an innovative smart service called the Federative Library (FedLib) that will be the first instantiation of the OOII. The chapter concludes with a summary and further research questions.

BACKGROUND

Several authors have considered the impact on business that comes with ubiquitous computing. Most work starts with a discussion of technological progress and the business impact arising from that progress. Roussos (2006) adopts this perspective. In his view, contingencies for ubiquitous commerce arise from emerging technologies which again have an impact on business innovation, as well as on society. Similar to that view, Mattern (2005) discusses technological progress that allows deploying smart objects in an informatized world around us. Within this world, informational counterparts of everyday objects and of people induce social and political challenges and pave the way to new applications. Fleisch and Tellkamp (2006) act on this assumption and provide a framework that helps to identify value-creating application areas and discuss challenges in implementing these applications. Whereas Fleisch and Tellkamp focus on processes within a company or between companies, Fano and Gershman (2002) consider the impact that smart services will have on relationships between customers and the service provider. These relationships will be profoundly affected by new classes of services emerging in smart environments. Based on the vision that a service provider will steadily be able to access customers, they make a set of assumptions about how services will be provided in a world of ubiquitous computing.

By discussing technology, smart environments and several business aspects, the present literature gives considerable insights into how ubiquitous computing will impact future business. However, current frameworks do not even consider the potentials for innovation. For example, the German Federal Office for Information Security provides a framework to discuss the socio-economic impact of ubiquitous computing (BSI, 2006). It considers “determining factors,” “enabling technologies,” “pervasive computing features,” and “Impacts” without addressing innovation. Therefore, we propose in the following section our framework for ambient business before discussing the impact of ambient business on innovation.

THE FRAMEWORK FOR AMBIENT BUSINESS

The framework for ambient business divides the analysis of ubiquitous computing in the context of entrepreneurial activities into four perspectives (see Figure 1).

First, the emerging technologies layer focuses on advances in technology during the last few decades that allow the development of new types of IT-infrastructures. These emerging technologies present new qualities that open up new forms of IT usage covered in the smart environments layer. Thus, new applications (smart services layer) can be implemented within smart environments. Finally, the ambient business layer deals with the impact of these innovative smart services on economy, business environment, and society.
Emerging Technologies

Advances in networking as well as embedded technology, display technology, and nanotechnology have been driving the research on ubiquitous computing during recent past decades (Hilty, 2005). Smaller and more powerful, but less power-consuming, information technology has facilitated the development not only of common IT-devices. Even objects such as coffee cups (Gellersen, 1999) and pens (Anoto, 2007) can be equipped with sensors, processors, storage capacity, and actuators. The objects can be connected with each other in order to “cooperate,” thus forming joint embedded IT environments. As a result, new opportunities for applications scenarios arise.

Current progress in the layer of emerging technologies is indeed quite remarkable. An in-depth discussion of technologies, therefore, is beyond the scope of this chapter. The remainder of this chapter assumes a portfolio of technologies which designers and developers of applications and services can use as a basis for presenting technical solutions for application demands in a world of ubiquitous computing. In this context, Semantic Web (Berners-Lee, 2001) technologies are of particularly great relevance.

Smart Environments

In the second layer, the qualities of emerging technologies are explored. The role of these qualities is twofold. First, they are the means for revealing new applications and services based on the emerging technologies. Second, when analyzing the shortcomings of current applications and services, the qualities can pave the way
Towards Ambient Business

for satisfying people’s innovation needs at the emerging technologies level.

An environment equipped with an IT-infrastructure as described above is able to gather information about people, things, and other surrounding conditions. The information can be used to recognize the context in which people and things are embedded, so that the environment becomes aware of the real world. Thus, the environment is able to estimate the needs arising from the real world and to adapt itself to (mobile) users and situations (Coutaz, 2005), or may even act autonomously. The environment becomes “smart.”

Emerging technologies not only facilitate extensive information gathering, but also provide new ways to interact with users. Information access and services presented by the smart environment can be almost seamlessly integrated into the user’s world. Interfaces become natural and interaction with them subliminal. The technology even disappears from the user’s perception (Weiser, 1991; Norman, 1998; Want, 2002).

Although different qualities reveal different potential applications, the qualities of emerging technologies are not independent from each other. In fact, like the emerging technology layer, only by analyzing the interplay of different qualities and by transferring them to the application domain, can the full breadth of application potentials be revealed. Accordingly, these qualities point to smart services that may be developed on top of smart environments.

Smart Services

Merging the real and the virtual world, every relevant object in the real world acquires an informational counterpart in the form of a data object (Mattern, 2005). Thus, a service that uses this informational counterpart may know the location and condition of a physical object. Such services are already applied to ensure efficient logistic processes (Angeles, 2005) and whenever assets have to been managed. Due to the decreasing costs of corresponding technology, services will not only be offered to companies, but a vast range of services is also conceivable for individuals. Books, clothes, and CDs may be managed in a similar manner in the future. By analyzing the current and past context of an object, which comprises additional information from outside the environment, appropriate supportive services can be developed and offered through a so called smart product (Maass, 2006). Additional services may be presented for use, for example, at a medicine cabinet when the medicine runs short (Floerkemeier, 2003). Such services can also be individualized, so that they adapt to the preferences, habits, and current situation of the user (Coutaz, 2005). When services become fully aware of their context, they could be invoked automatically, without explicit manual interaction by the user (Tennenhouse, 2000).

Ambient Business

The ability of smart services to adapt themselves to situations and provide a huge amount of information instantly not only supports current business processes and distribution channels, but it also opens new ways of doing business.

Services in the application domain of supply chain and asset management reveal opportunities for monitoring tangible assets more closely and thus creating a higher degree of control. Yet, the costs of control decline at the same time. Not only high-value products, but also lower-value ones can be managed by efficiently smart services. This results in a very tight mesh of control throughout the company (Fleisch, 2005). Through means of such control, processes can be adjusted to incidents that are internal or external to the processes. This may result in more efficient and reliable processes within an organization.

Inter-organizational processes that utilize smart services may also cut costs in the supply chain. Through crossing internal company bor-
Towards Ambient Business

ders, information flows that are highly aligned with the flow of physical goods reduce losses, and control overhead that frequently occur at such borders (Strasser, 2005). Furthermore by, opening up their data base, companies are able to offer smart services to each other, for example, control and information services (Fleisch, 2005). Functions that were previously closely linked to one’s own organization can now be outsourced. It becomes feasible for companies to offer their outstanding capabilities that were previously bound to the organization to suppliers, customers, and even to competitors.

In the future, products may be provided more as a bundle of the physical product itself and the services linked to it. It may be sold as a platform for generating additional revenues by means of services in the after-sale phase. In this case, the vendor comes closer to the customer through two means. First, he is able to address a customer not only through traditional sales channels like a store. He can also address the customer at the preferred location (Fano, 2002), which may have an enormous impact on business models and sales channels. Second, information may flow in the other direction as well. Users can provide usage information to the vendor and receive special services or a monetary reward in return. The vendor can use this information to align R&D-efforts to users’ needs. Going one step further, smart services can be used to involve users in the R&D-process, whether directly or via mediators like Innocentive (Innocentive, 2007).

It can be assumed that ambient business entrepreneurs will find themselves in markets with short innovation cycles. Like the internet, smart environments follow the end-to-end principle. This principle states that the intelligence of an information system rests, that is, in the end-user-devices, not in the network (Saltzer, 1984). Someone who can exercise control as a central authority typically owns the network. In contrast, the “ends” of the network are typically in user’s control. The user is able to replace or enhance the ends and to deploy innovative applications and content. Thus, Lessig (2002) argues, systems that put the end-to-end principle into practice encourage innovation. The short innovation cycles in the internet can largely be explained by the fact that they implement this principle. Smart environments are largely considered to be built out of smart devices and, therefore, follow the end-to-end principle as well. Users of smart environments can decide what device they want to use and what smart services they should deliver. Therefore, the innovation cycles in ambient business are likely to be at least as short as in Electronic Business. Moreover, consideration must be given to the fact that smart services can be invented around all instances of our life that allow the creation of manifold potential applications. New entrants may attend to these potential applications, enter existing markets, and shift competition (Watson, 2002). Companies have to reconsider how to cope with new competitors in a high-velocity market as well as what strategies and business models to apply.

Smart environments and services that are deployed in our everyday life may have considerable social, ethical, and legal implications which also have to be considered in the business perspective. Different studies reveal corresponding potentials and perils. Positive effects can be expected throughout all activities of daily life. In particular, elderly people and handicapped persons may benefit from supportive applications and natural interfaces (Coroama, 2003). These interfaces may also close the digital divide between those with and those without access to computing facilities (BSI, 2006). Hilty et al. (2005) raises the question whether positive effects resulting from efficiency enhancements may be partially undone by side effects, that is, by increasing people’s stress level, and whether smart products may lead to pollution and health hazards. However, most concerns relate to the amount of gathered data and resulting privacy issues. History suggests that balancing companies’ benefits and customers’ privacy is
Towards Ambient Business

crucial for the success of commercial smart environments (McGinity, 2004). Bohn (2005), Hilty (2005), Bizer (2006), and BSI (2006) provide an in-depth analysis of possible perils and give hints how to balance benefits and user’s privacy. Lahlou (2003) proposes a set of privacy guidelines to cope with this challenge.

In addition to these social, ethical, and legal implications, the aim of this chapter is primarily to discuss the impact of open innovation in a world of ambient business from an economic perspective.

OPEN INNOVATION FROM AN ECONOMIC PERSPECTIVE

The idea of encouraging innovation throughout openness with regard to IS was widely discussed in the early 1980s when mainframe vendors started to restrict the use of object program and source code of formerly unrestricted software. As a countermove and in order to gain independence of the software industry, Richard Stallman (1999) propagated the idea of free software. This paradigm shift away from the conventional IP regime resulted in software that can be used, modified, and redistributed by everyone. Related movements like Open Source (Lerner, 2002; O’Reilly, 1999), Open Standards (Perens, 2006; West, 2004), Open Systems (Gabel, 1987), and Open Content (Ciffoilli, 2003; Wiley, 1999) all follow a similar idea: By modifying and recombining existing work, innovation can be collectively achieved. Thus, self-organized ecosystems of Open Innovation emerge when independent actors contribute to a product system (the product and its complementary assets). These Open Innovation ecosystems have proven that they are able to compete with companies that are reluctant to open up their systems and their IP. Software like Apache, Mozilla, and Linux has achieved a considerable market share, and Wikipedia comes close to the quality of the Encyclopaedia Britannica (Wales, 2005).

The idea of openness may contradict the common approach of value extraction of a kind of assets, that is, intellectual assets, in most societies. It is widely assumed that these assets can and should be owned by someone and hence are the IP of the creator or perhaps of the buyer. However, such IP shares the fate of most intangible assets. Once it is revealed and brought into the public domain, there is the risk that it becomes common. The creator and his competitors can use the asset to create value. The potential for the creator to differentiate his products from those of competitors is reduced. Hence, the use of the asset by competitors reduces the creator’s return on the asset (Magee, 1977). In order to ensure compensation for the creators of innovative intellectual work, legislators enact laws to protect intellectual property rights (IPR). Creators and buyers of special rights are exclusively allowed to utilize property for a certain period of time. Through these laws, intellectual work should be rewarded and innovation stimulated (Chesbrough, 2006).

The management of intellectual assets often focuses on closing up companies’ borders from the intellectual assets point of view in order to keep others from harvesting the fruit of their labor. However, legal instruments often cannot hinder third parties from gaining large shares of the profit arising from innovations: The protection mechanisms are often not enforceable. They are easy to bypass by using slight modifications of the IP in question, and this results in various forms of uncompensated adaptation (Teece, 1986). Under certain conditions, IPR can be ignored. The music and software industries are examples of this phenomenon. Thus, IPR are often not capable of protecting the owner from the exploitation of assets by others.

Closed companies’ borders may not only frustrate the aim to protect assets; in fact, they may prevent companies from gaining profit from their assets at all. R&D departments regularly create intellectual assets that cannot be utilized within the company. Procter & Gamble, for example,
estimated in 2002 that it only commercialized 10% of its patents, whereas the remaining 90% virtually gathered dust on shelves (Sakkab, 2002). Licensing may be a viable option to exploit unutilized patents (Rivette, 2000). However, potential licensees are often reluctant to adopt external technology (Chesbrough, 2006), and the legal protection of an IP is often too weak to make licensing a viable option.

Companies such as Google and Amazon turn in part away from the closed way of managing intellectual assets and appear to profit from opening up their information systems and, therefore, revealing some of their intellectual assets. As another example, IBM donated 500 patents to the Open Source community in 2005 instead of generating revenue by licensing it. Moreover, IBM spends $100 million each year in the development of Linux. IBM opens up these assets to regain a prominent role in the operating system market—a position it had lost beforehand with AIX (Chesbrough, 2006)—and is now able to use Linux as a platform to sell products and services around Linux (West, 2003).

**Drivers for Open Innovation in Ambient Business**

To reveal the drivers for making assets publicly available in a world of ubiquitous computing and the resulting potential benefits, the open approach can be discussed from a resource perspective (Barey, 1991; Penrose 1959; Wernerfelt 1984). This perspective reveals the possibilities arising from a resource exchange at the layers of the device infrastructure, the information infrastructure, and of the smart service.

Resources can be considered as tangible and intangible assets that are tied somehow semi-permanently to the firm (Wernerfelt, 1984). They compass all input factors that are needed to create value during a product life cycle. Companies can leverage the resource base to achieve a competitive advantage if these are valuable, rare, imperfectly imitable, and nonsubstitutable (Barney, 1991). As such, innovative related resources that fulfill these requirements can contribute to achieve an advantageous position in markets in which innovative products and services are valued by customers.

As argued before, in a world of ubiquitous computing, innovation life cycles may be short. From innovations in technologies and services often spring a disruptive force that leads to changes in the market environment: New innovators entering the market and unpredictable demand shifts may change the market structure. However, competitive advantage can only be achieved when the company’s resources are in line with market conditions (Eisenhardt, 2000). Yet, in the changing markets of ubiquitous computing, frequent adaptation of the resource base is necessary. On the one hand, this is a major challenge for companies in UC markets. On the other hand, this should be seen as an outstanding opportunity to gain competitive advantage through superior capabilities in resource acquisition, namely adaptation, integration and reconfiguration of internal and external resources (Teece, 1997).

However, companies may not be able to alter their resource base on their own fast enough, and necessary resources may not be available for licensing. For example, when Microsoft released Windows NT and Windows 95, Apple had to innovate in its operating system to cope with Microsoft. However, Apple was not able to alter its resource base accordingly. It, therefore, published its source code partly as an open source and allowed external programmers to contribute to the software development process (West, 2003). Following this example, opening up a company’s border and letting resources and, thus, intellectual assets pass through may be a viable strategy for acquiring new resources.
Towards Ambient Business

**Open Innovation in Ambient Business Settings**

In ambient business, diverse resources can be incorporated to enhance the innovative potential that is associated with a company’s innovation processes and the innovation potential of its products. These resources can be obtained from all stakeholders of the company’s products, especially customers, suppliers and manufacturers of complementary products, and also from competitors as well. Accordingly, a transfer can be conducted at every level of the proposed framework.

The device infrastructure builds up a smart environment and, hence, provides the foundation for smart services by gathering data and interacting with the user. Access to a device infrastructure is an essential asset for service providers. A tightly restricted and exclusive access to the IT-infrastructure through proprietary standards allows the creation of a monopoly position. This facilitates the monopolist to set rules which rivals must obey (Adams, 1982). Furthermore, the proprietary standards of the infrastructure create a barrier for new entrants because a change to another infrastructure is linked with incompatibility costs. However, providers of complementary devices and services are likely to be discouraged from using a tightly controlled IT-infrastructure (Farrell, 1985; Garud, 2002). Therefore, it may be impossible for a company that is playing the proprietary game to build up the full range of different devices of various application domains. Hence, if a company opens up its device infrastructure and relies on open standards, access to foreign innovative components may be gained by allowing them to become integrated into the company’s own infrastructure.

Garud and Kumaraswamy (1993) showed through the Sun Microsystems case how companies could improve their strategic position by implementing an open system strategy. By relying on open standards and by liberally licensing some of its own know-how, Sun attracted capable complementors that helped to improve the overall systems performance. Sun ensured its head start in competition by conducting shorter innovation cycles than its competitors.

Another strategy to ensure returns is to retreat to a certain layers of the product architecture. Layers that do not contribute to a distinctive resource base generally do not help companies to gain a competitive advantage. By opening up these layers, an unrestricted foundation for higher layers may be laid that will be further developed by external innovators. Thus, the company may concentrate on extending its distinctive resource base on an upper layer, such as developing smart services. Apple, for example, published only source code of the operation system’s lower levels. However, the Graphical User Interface and certain APIs, with which Apple can differentiate its operation system from its rivals, remained closed source (West, 2003).

As demonstrated earlier, the information infrastructure is the basis on which services are growing. It is, therefore, an important resource. The richer the information pool in terms of quality and quantity, the more services can be provided. These services will also be more reliable. For that infrastructure, the same principles apply as to the device infrastructure, since both are infrastructure layers. The information infrastructure that is available can be enhanced by opening it up and allowing other information infrastructures to interlink with it. Furthermore, companies may benefit from receding that layer and to commercialize the smart service layer. As distinct from the device infrastructure, when considering the information infrastructure, the user may play a more important role. Sensors deployed in the environment allow gathering data from a user’s context without bothering the user, who may herefore willingly contribute this information to a company’s information infrastructure assuming he trusts it. In this case, a services provider as
Towards Ambient Business

well as a user may benefit from highly customized services. The user may play a more active role when it comes to semantically enriched data. Such enrichment can only be automatically conducted in exceptional cases.

Such tasks commonly require human judgment, for instance, to decide whether the Eiffel tower is ugly or beautiful. By opening up the base data, users can be encouraged to fulfill the tasks. Amazon’s mechanical Turk is an example for a platform that brings together tasks that need human judgment to be solved and people that are willing to solve such tasks. Persons like students, housewives/husbands, and pensioners may be attracted by being paid a small amount of money (Amazon, 2007). A voluntary participation, however, may be provoked whenever the user can benefit from an enriched information pool or consider his participation as an act of self-expression. Google Earth is a prime example of users being willing to contribute to a (partly) opened database (Roush, 2005). Other motivation mechanisms exist, such a mechanism, in which the work is disguised as a game (Ahn, 2006).

One of the major challenges of innovation—and therefore of creating smart services—is bringing together need information and solution information. These two types of information cannot usually be found at one and the same site (Hippel, 1994; Thomke, 2002). As a resource, solution information encompasses all information that provides a company with the technological competence and the potential that is necessary to supply customer needs and that resides traditionally in the company’s site (Thomke, 2002; Thomke, 2003). Need information encompasses the user’s preferences and requirements with respect to new products and services (Reichwald, 2006). It is generally in the user’s domain. The innovator’s lack of need information is a critical drawback for creating smart services. Thackara (2001) states that in the current ubiquitous computing research, developers know how to create amazing things with ubiquitous computing technology, but they lack in need information. They do not know how to use their knowledge to create solutions that benefit the user. However, obtaining this information is far from easy. Need information is somewhat “sticky,” that is, it can only be conventionally revealed from potential users through costly and time consuming methods. Users often do not fully understand their needs or are not able to explicate them (Hippel, 1994). Smart environments can help to overcome this problem by studying contextual data gained from the user. Thus, it may be able to observe how a user behaves in using a smart service, which problems he encounters and what needs may remain unfulfilled.

Another way to overcome the problem is suggested by von Hippel (Hippel, 2002, 2005). By outsourcing need-related tasks to users, the problem of how to directly acquire sticky need information can be avoided. Thus, the user handles a task, which the company formally conducted, and returns a solution he is satisfied with. This solution is commonly less sticky than the need information itself. However, to fulfill the task, the user needs relevant solution-information that sticks with the company. Here, the company partly opens up its information pool by providing a design tool, called toolkit, to the user. This toolkit implements the company’s solution information, which is provided in such a manner that the user can handle it easily. Smart Environment may broaden the facility of toolkits. Sophisticated toolkits may be able to be provided to the user at that point where his need generally occurs. Multimodal interfaces (Dam, 2001) may facilitate a richer interaction with the toolkit and, thus, allow solving complex tasks. Tangible interfaces can be used to let the user not only interact by “mediating” devices such as keyboard, mouse, and display, but to directly manipulate physical objects (Sharlin, 2004). An example of such a tangible interface is the “illuminating clay” from Piper et al. (2002). The computing system is able to gather the form of the clay the user works with and is able to augment it with additional informa-
Towards Ambient Business

The embedding of computers or smallest processors, memory chips, and sensors into physical objects, referred to as smart objects, constitutes a step toward connecting the real with the virtual world. Other examples are sensor networks—smallest sensors introduced into the environment, forming ad-hoc networks in order to recognize, for example, the whereabouts, speed, size, and form of objects and to observe them over a period of time (Culler, 2004). Sensors may perform an important role in the context of managing object information. This is particularly true for the use of sensors which permit a clear identification of smart objects by remote enquiry. With the assistance of Radio Frequency Identification (RFID), objects are identified automatically without line of sight, by transponders which are attached to them (Finkenzeller, 2003).

It has to be taken into account that the information content of sensor data is rather limited. Hence, the data must be gathered and formed into usable events by means of appropriate filter and aggregation procedures. Numerous projects for the development of suitable middleware have already commenced, but focus primarily on integration of sensor data in closed, business-related applications or communication between smart objects (for an overview of selected projects, see Henrickson, 2006 and Schoch, 2005). Most projects, however, ignore the further use of the information in other contexts or by external developers.

Ensuring the utilization of gathered data for external developers requires extending existing middleware technologies with mechanisms for the semantic markup of sensor data. Semantic Web standards such as Topic Maps (Durusau, 2006; Garshol, 2006; Pepper, 2001;), or RDF(S)/OWL...
Figure 2. The open object information infrastructure (OOII)
Towards Ambient Business

(Brickley, 2004; McGuinness, 2004) can be employed for this purpose. Several middleware technologies have already been extended with mechanisms for the semantic markup of sensor data. SOAM (Smart Object Awareness and Adaptation Model) (Vazquez, 2006) is one initial approach. SOAM is a smart environment reactivity model in which user preferences are formalized using the Standard Ontology for Ubiquitous and Pervasive Applications (SOUPA) (Chen, 2005). These preferences lead to adaptation on smart objects without explicit user intervention, because of the use of automatic, semantic reasoning with respect to the environmental conditions.

An extended version of this ontology is also used by the Context Broker Architecture (CoBrA) (Chen, 2005). This is a broker-centric agent architecture for supporting context-aware systems in smart meeting rooms. The project goal is to create a meeting room that can facilitate typical activities as setting up presentations (allowing users to control services via speech), or adjusting lighting and background music in a room in a manner that is tailored for the specific meeting.

Even if these projects incorporate semantic technologies, they do not consider the reuse of their smart environments within external smart services. Nevertheless, they could be suitable for application within the OOII, as they enable the combination of ubiquitously gathered information with additional semantically-enriched information provided, for example, by the Semantic Web.

Building Up Open Information

In developing smart environments, the large number of smart objects must be taken into account. Additionally, the prevailing energy and capacity limitations of embedded computers and sensors make it necessary to save accompanying virtual data objects and to process information on additional network nodes which have sufficient memory capacity and computer power. These prerequisites advise the implementation of peer-to-peer (P2P) networks that offer better scalability and a self-organized and decentralized coordination of unused or limited resources, compared to the alternative client/server architecture (Barkai, 2001; Milojicic, 2002; Oram, 2001; Schoder 2003). Other characteristics, including a higher fault tolerance and a more effective support of spontaneous networking of entities, may be advantageous as well, because no continuous connection can be ensured for smart objects, due to their mobility.

In order to enable Open Innovation, arbitrary access to the stored semantically enriched object information must be ensured for external developers. As the Napster case revealed, central approaches tend to limit access to the information offered or even deny it completely. However, P2P projects like Freenet (Clarke, 2000) and Free Haven (Dingledine, 2000) have demonstrated how an anonymous and censor-free information access can be made possible. Using P2P networks for storing and retrieving semantically enriched information requires an extension of possibilities to pose semantic queries. On the one hand, scientific research is pursuing scheme-based approaches, and, on the other hand, ontologies or P2P Semantic Web services are applied (for an overview of selected projects, see Staab, 2006).

Using semantic query facilities, various developers with different perceptions can search for relevant (object) information, combine the information freely, and, hence, create innovative services (as indicated in Figure 2). For example, information about the speed of vehicles (low speed for a number of cars) combined with their actual location (they are all in a certain area) and additional information (a road map) might indicate that they are stuck in a traffic jam on a highway. This information could then be provided to others as a traffic-information service. In order to facilitate even the reuse of this new smart service, the service has to be specified semantically and published in the OOII. Subsequently, other developers can reuse this service (and the con-
Towards Ambient Business

connected smart environments and objects), combine services with information gathered by another local sensor network (the outdoor temperature in a certain area is below zero), and set up a warning system (you are traveling to a snowy area—do not forget your snow chains).

In order to discover and integrate semantically enriched information, a number of semantic web services (Alesso, 2004; Fensel, 2002) and smart agents (Konnerth, 2006; Lee, 2005) have been developed. However, these projects mainly solve negotiation problems with respect to a common understanding of certain application contexts. The user-oriented framework of Task Computing (Masuoka, 2004) is a promising approach in the context of ubiquitous computing. It uses the Task Computing Environment Object Ontology (TCE, 2004) and is designed to operate dynamic ubiquitous environments in which a mobile computing user dynamically discovers the current set of available semantically defined services (Song, 2004). The end-user can easily and seamlessly integrate and manipulate services found on their own computer, the nearby devices and relevant remote Web services. However, Ben Mokhtar et al. (2006) claim that even this opportunity, or need to select the right composition among all the possible compositions suggested to the user, presents a major drawback of this framework. Accordingly, they developed a COnversation-based service Composition middleware (COCOA) that minimizes the necessary user interaction by enabling the automatic and transparent deployment and execution of user tasks.

Nevertheless, these approaches are limited to the discovery of smart services within smart environments in order to improve user interaction. They do not address global service discovery as needed within our OOI. METEO-S WDSI (Verma, 2005) and GloServ (Arabshian, 2005) are approaches that perform global service discovery through the use of P2P networks. METEO-S WDSI presents a scalable environment for Web service publication and discovery among multiple registries in general. It uses an ontology-based approach to organize registries into domains and enable domain based classification of all Web services. Each of these registries supports semantic publication and the discovery of Web services. GloServ is a global service discovery architecture that can be used to build a distributed context-aware service discovery system related to pervasive computing scenarios (Arabshian, 2006). Currently, we examine which one of these approaches would be applicable within the OOI.

Building Up Smart Services

Based on Open Information, innovative services can be developed. The Federative Library (FedLib) represents one first instantiation of our OOI (see Figure 3) and demonstrates the feasibility of the design. The FedLib is a library service for work groups that functions without the need for collecting books at a centralized location. In fact, the collection is held by the participants themselves. Book owners decide about making their books available for use by others. Thus, originally exclusive private collections of books that are common in organizations could be opened up to other would-be readers.

The pilot implementation of the FedLib will be realized within the Department of Information Systems and Information Management and includes our own office rooms and books, as well as books from the faculty library. We already built up the necessary smart environment and started to extend a reference information manager to the corresponding search-client in order to integrate the smart service into the daily workflow of a research assistant. Besides building up the smart environment, we focus on reviewing existing technology for further use within the FedLib and OOI.

Implementing the smart environment of the FedLib, bookshelves and doors of our office rooms are equipped with RFID readers (see Figure 4). Books labeled with RFID tags are distributed in the
Towards Ambient Business

Figure 3. The Federative Library (FedLib)
Towards Ambient Business

participants’ offices. In order to enable ubiquitous information gathering about the presence of books, the detection of a book’s current location must be combined with former “reads” to corresponding events. For instance, when the system detects a book at the door reader in Office 1, information must be available as to whether the book is leaving or entering the office. It is leaving if the book was previously detected on the bookshelf in Office 1. It is entering if the book was previously located at the door reader in Office 2.

By aggregating these semantically enriched events with additional information about the occupant of each office, the FedLib can track and allocate books automatically in terms of the current borrower. This minimizes administrative overhead, due to the ubiquitous integration of processes in the daily workflow of the research assistants. It therefore becomes possible to borrow books by collecting them directly from a colleague’s shelf, taking them to one’s own office, and placing them at the table there. In order to return a book, it can simply be put back on one’s own bookshelf. During that natural interaction with the Federative Library, the location and status of the book changes according to the events occurring (numbered 1 to 5 in Figure 4). In order to integrate even the management and the retrieval of books into the daily work of a research assistant, the corresponding search-client is integrated in a reference information manager. Furthermore, services like Amazon’s Web Service (Levitt, 2005) and special semantically-annotated resources produced by Semblogging (Cayzer, 2004), are suitable for integration into the manager in order to provide additional useful information.

CONCLUSION

Previous research and the movement towards openness suggest that an opening of information systems may reveal new contingencies for innovation. In order to analyze these contingencies in a world of ubiquitous computing, we applied the framework of ambient business. We showed that
Opening up smart environments as well as smart services, reveals potentials for innovative services that can be created by external developers. Thus, the innovative strength of companies that act in open systems cannot be achieved by those that rely on rather proprietary information systems. However, most contemporary projects ignore the further use of the information in other contexts or by external developers.

By designing the Open Object Information Infrastructure, we showed how to open up ubiquitous computing systems. Within this infrastructure, information pools that emerge from smart devices are connected by existing Semantic Web standards. This allows developers to search for and combine information they need, create innovative services, and reuse and combine them in an innovative manner. By introducing the first instantiation of the OOII, we demonstrated the feasibility of our design. The Federated Library combines information from decentralized objects and external services to an innovative library services that cannot be realized in closed environments.

However, the open approach is not only applicable from the technological point of view. Economic rationality can argue similarly in favor of opening information systems and thus companies’ intangible assets. In some instances, property rights cannot fulfill their aim of fostering innovation, but in fact encumber a further development that builds up on the protected assets. Therefore, in terms of economic and social welfare, the guarantee to freely use innovation as well as the incentive to commercialize innovations has to be perfectly balanced.

Additionally, a cross-border flow of intellectual resources can be beneficial. In high-velocity markets as can be expected in ambient business, the alignment of company resources to the corresponding market becomes critical. In particular, innovation-related resources have to cope with market developments. Opening up the company’s borders can be a strategy for acquiring external innovation-related resources and, therefore, a head start in competition with closed innovators.

**FUTURE RESEARCH DIRECTIONS**

Questions arise regarding the willingness of market players to adopt the idea of Open Innovation. First, open smart environments may infringe privacy. Consequently, we may ask what information a user considers private and in what situations. What information should, therefore, not be freely available to service providers? How should laws be formulated in a world in which information is freely exchanged? The first studies have already appeared which analyze the dangers associated with the emergence of ubiquitous computing (Hilty, 2005; Bizer, 2006; BSI, 2006), but they are only the first steps towards answering the core questions.

Likewise, companies must consider which resources and information should be allowed to cross their borders. Sophisticated strategies have to be implemented to ensure the appropriability and, therefore, the return on the innovations. Thus, a manager may ask what products and resources should be opened and to what extent. What are the company’s distinct resources that have to be maintained inside the company? What kind of resources can be utilized more effectively outside? Control of the resource transfer must be established.

Furthermore, organizational structures and processes have to be adapted to facilitate the adaptation of external resources, and employees have to be trained to interact with external stakeholders. Moreover, external innovators have to be attracted. When increasingly more companies try to harness users’ resources and creative power, how will the users react? Will their attention and time become a limited resource so that companies have to compete for it?
REFERENCES


Towards Ambient Business


Fab@Home (2007). Fab@Home. Retrieved on April 10, 2007 from http://www.fabathome.org/


Towards Ambient Business


**ADDITIONAL READING**


Chapter 6.15
Successes and Failures of SAP Implementation: A Learning Perspective

Tanya Bondarouk
University of Twente, The Netherlands

Maarten van Riemsdijk
University of Twente, The Netherlands

ABSTRACT

In this article, we conceptualize the implementation process associated with SAP_HR as an experiential learning one (Kolb, 1984), and analyze qualitative data collected during discourse analysis during a six-month case study. We saw that a lack of communication plus misunderstandings between the different parties involved in the project led to mistakes in working with the system. However, with encouragement from the “top” to improve learning, working with the system became easier for the whole group involved and for the individual users. Although Kolb’s theory is widely acknowledged by academics as a fundamental concept that contributes towards our understanding of human behavior, we propose another use: to consider this theory in association with an IT implementation strategy to identify the mechanism of IT adoption in an organization.

INTRODUCTION

This article is about the implementation of information technology (IT). Although this topic has been debated for several decades, the practical arena is still confronted with the so-called “go-live” problems with IT projects being less successful than predicted.

It is widely recognized that “go-live” IT use often develops in a different way than that in the plans made, and that the degree to which the use of a technology corresponds to the anticipated rules and norms can vary considerably, depending on the organizational context, the type of
IT, and the end users’ awareness of the system. (Arkich, 1992; Bardram, 1998; Bikson & Eveland, 1996; Ciborra, 1996; DeSanctis & Poole, 1994; Orlikowski, 1996).

There has been much research with consequent recommendations on how to introduce a new IT to employees in order to minimize or avoid troubles during IT projects. However, IT introductions are still known to be time consuming, indirect, and sometimes impulsive developments, leading to a mismatch between the initial ideas behind the information technology, and its real use seen through the employees’ perceptions and experiences.

So, why have we apparently learned so little from IT failures? There are two popular reasons. It is argued that traditional IT research, focused as it is on IT implementation factors, has failed to identify the true causes of failure (Sauer, 1999). Perhaps, the traditional factors and processes considered are symptoms of, rather than the reasons for, the failure and, if so, attacking the symptoms will not cure the disease. The other popular possible reason is that even if the identified failure factors are the causes of the IT failure, they are not easy to avoid (Kling, 1987). Arguably there is a third possibility: that the traditional studies do not mirror the interactive, complex reality of the IT implementation process in which users are involved.

Given this situation, this article proposes to look at IT implementation from an interactive prospective, focusing on the collaborations among users while working with the newly introduced technology. Almost all modern technologies have networked, or collaborative, fragments, and users are engaged in common tasks through these “fragments.” With the rise of wireless, mobile, and Internet technologies on the one hand, and with integrated office environments on the other, organizations are increasing their demands for cooperative working. Cooperative “fragments” can often be recognized in various work situations ranging from document sharing, cross-functional and cross-departmental projects, to incidental correspondence between employees linked by a given task. Stand-alone computers nowadays are generally used only for trials and experiments in organizations, the most common situation sees workstations hooked up to an organizational network.

Fundamentally, users communicate with one another when using IT. The communications during an IT implementation project reflect a situation in which groups of users are developing a common understanding of the technology they are forced to (or want to) use through learning processes amongst themselves.

The importance of several aspects of learning within collaborative settings has been seen in various IT studies:

- Changes in technology may lead to changes in various aspects of professional competency such as knowledge, skills, and attitudes. These, in turn, can influence the ongoing use of the system. Hence, in theory, there is an ongoing process of professional and technological development which is referred to as a learning process by Neilson (1997).
- User groups have to adapt to a novel way of working when a new technology is introduced. Adaptive structuration theory has shown that not all groups do this in the same manner, and the adoption process, referred to as “appropriation,” depends on the group processes and the way in which people interact with one another (DeSanctis & Poole, 1994; Hettinga, 2002; Ruel, 2001).
- In the “extended version” of the structural perspective, Orlikowski (2000) proposes looking at “communication, mutual coordination, and storytelling” as important determining sources for engagement with the system (p.411).

Although some “feeling” for the topic now exists, and recent research has emphasized the
importance of certain elements of learning for IT implementation, systematic insights are still lacking. In this article, we propose a new view on IT implementation; one that considers the learning processes as the key factor in “getting used” to a new system. To the end, we strive to answer the question as to the role of group learning in the IT implementation process, and to illustrate our answer with findings from a longitudinal case study.

In the following sections, we first briefly present the basics of learning, and especially of experiential learning, in IT implementation. We then discuss the research methodology, followed by the empirical results from a longitudinal case study in a Dutch university.

LEARNING AND IT IMPLEMENTATION

An examination of existing studies has revealed that there is little theoretical or empirical research with a particular focus on the role of learning in explaining and resolving the problems of implementing and using information technologies in organizations. The emerging studies that do attempt to address both organizational learning and information technology consider learning as an alternative antidote to the organizational struggles associated with IT (Robey, Boudreau & Rose, 2000).

The current view of the link between “learning” and “IT implementation” in the literature can be seen as two-fold: as formal training as a way to overcome knowledge barriers, and through the role of experience in IT implementation.

Literature on formal training in IT implementation usually focuses on the individual level and deals directly or indirectly with overcoming barriers to acquiring new knowledge in IT use.

The second, and the major, literature stream on learning and IT implementation results from research on experience-based organizational learning. There is strong evidence suggesting that an organization’s own experiences provide a knowledge base which guides future actions. Case study literature provides greater detail on the role of experience in IT implementation: some offer evidence of the benefits of experience in achieving a successful implementation (Caron, Jarvenpaa & Stoddard, 1994; Yetton, Johnston & Craig, 1994); while others illustrate the difficulties of learning from experience (Ang, Thong & Yap, 1997; Robey & Newman, 1996).

However, the reported studies do not discuss instances where organizations fail to learn from their own experiences. Another limitation is that the authors do not discuss the “competition” between recent and earlier events as sources of experience. How can an organization adapt an old experience to a new situation? Clearly, learning from experience is more complex than simply adjusting current actions based on previous outcomes. What are the common key issues and processes in experience-based organizational learning? How can one transfer conclusions from an IT experience in one company to another, and is this even necessary? Finally, when and where are the lessons applied and really learned? Such questions remain unanswered in the existing studies.

Organizational Learning

Before introducing our concept, we briefly discuss the main issues associated with organizational learning. This is presented in the literature as a combination of two different approaches: learning as an outcome or “intended product,” and learning as a process. For example, Agyris and Schön (1978) define learning as a process of detecting and correcting error. Kolb (1984) stresses the importance of the transformation of human experience (the process) that leads to new knowledge (the result). It is also recognized that an outcome of the learning process could be a more experienced person, one who might have
a changed self-concept (Jarvis, 1987). Marsick (1987) focuses on learning as acquisition, interpretation, or assimilation of information, skills, and feelings. The learning-as-object approach is still influential. However, of late, there has been an increasing emphasis on learning-as-process (Gourlay, 2006; Tsoukas & Vladimirou, 2001; Walsham, 2005). Learning is increasingly being seen as a process of drawing “distinctions within a collective domain of action, based on appreciation of context or theory” (Tsoukas and Vladimirou, 2001, p. 979). Learning is further considered to take place only within communities, where human performance is articulated through social interaction (Walsham, 2005, p. 10).

This article focuses on the dynamic acting-interpreting learning that is “deeply involved in human processes of communication, and which cannot be divorced from the context” (Walsham, 2005, p. 7).

In this study, in an attempt to bring learning into IT implementation, we define learning as all the interactional processes through which users develop their understanding of a newly introduced technology and that helps them to implement it.

**Experience-Based Learning**

Within learning-as-process concepts, a further refinement is made in the direction of experiential learning (Kolb, 1984) in which the concept of “experience” is central. This provides an opportunity to “begin” the learning process only after the employees get a new technology and start working with it; that is when they gain “experience.”

Learning involves the interplay between two interdependent dimensions of knowledge: acquisition and transformation. Knowledge acquisition requires an individual to resolve the tension between apprehension (concrete experience) and comprehension (abstract conceptualization). Apprehension requires an individual to accept new knowledge through direct experience with the world (feelings and emotions). In contrast, comprehension occurs when an individual obtains knowledge through abstract concepts; in other words, when a person breaks down experience into meaningful events.

Transformation is another dimension of knowledge with a dialectical tension: between intention (reflective observation) and extension (active experimentation). During knowledge intention, a person learns by reflecting upon previously acquired knowledge. In contrast, learning by extension requires an individual to interact with an external environment.

In responding to the dialectical tensions of knowledge, individuals orchestrate their way around the cycle in a continuous process of interactions between personal and environmental demands (Kayes, 2002).

The learning cycle begins when one experiences the world through one’s actions of “doing” (Dixon, 1994). This immediate, concrete experience is the basis for the next stage of “reflecting,” which allows us to learn from our experiences. Observations are assimilated into a theory that makes sense of what we have experienced. The third step is abstract conceptualization, or “thinking.” The final phase is “deciding,” and actively testing the concepts that have been created from a real world experience. Following the fourth step, new implications leading to concrete action can be developed. Thus, one continually cycles through a process of collecting experiences, or a set of conceptualizations (Swieringa & Wierdsma, 1994).

Although there are other models of experiential learning, Kolb’s theory continues to attract attention because of its completeness and generalizability. Since 1971, over 1,500 studies, refereed articles, dissertations, and papers have reflected the work of Kolb, and provided insights into a broad range of learning processes (Kayes, 2002). The basic “wheel” has appeared

Kayes (2002) noted that Kolb had outlined the relationship between individual and social knowledge:

Apprehension of experience is a personal subjective process that cannot be known by others except by the communication to them of the comprehensions that we use to describe our immediate experience... From this it follows that there are two kinds of knowledge: personal knowledge, the combination of my direct apprehensions of experience and comprehensions I use to explain this experience; and social knowledge, the independent, socially and culturally transmitted network of words, symbols and images that is based solely on comprehension. (Kolb, 1984, p.105)

If we understand learning as changing knowledge and behavior through actions with IT, then we should acknowledge the importance of social experience and context in learning (Barrett, Cappleman, Shob & Walsham, 2004; Thompson & Walsham, 2004; Tsoukas & Vladimirou, 2001), and therefore also the importance of interactions among individuals.

**IT Implementation as Experiential Learning**

Although Kolb’s theory is widely acknowledged by academics as a fundamental concept that contributes toward our understanding of human behavior, we propose another view: to see Kolb’s theory as reflecting a kind of IT implementation strategy and suggesting a mechanism for IT adoption in organizations.

Learning begins with the experiences and actions of the targeted employees when a new IT is introduced. It is likely they will have to start operating with the system in order to execute their tasks. This process can develop from operating with the basic system modules in performing their everyday tasks through to searching for new techniques and possibilities present in the system (West, 2000).

The next activity in the learning cycle is reflection and this can occur at various points: after carrying out a few operations with the system, later during implementation, but also even before the system’s introduction as future users discuss technology design issues. Discussions, open dialogue, focus groups, and meetings with a project team might well focus on speaking out about difficulties or perceived difficulties in using the system. Users might express doubts and suspicions, or trust and belief in the existing ways of solving IT-related difficulties, consider the possible reasons for, and the outcomes of, mistakes made when operating the system, or discuss errors that occur in working with various IT functionalities (Schippers, 2003; Stahl, 2000; Tucker, Edmondson & Spear, 2001).

The conceptualization process may also take on a great variety of forms, but will always lead to a shared interpretation of the system among the users (Mulder, Swaak & Kessels, 2002). They will share their understandings of the role of IT in the company and its intentions towards every member of the group, as well as the design.
intentions of the system’s developers (Nelson & Cooprider, 1996; Hendriks, 1999).

Planning, when initiated by the users, aims to establish agreements on improving the use of the system. Employees may take the initiative to arrange additional training sessions, develop further instructions or manuals, or carry out other related activities. Developing policies in order to improve the use of a technology can become a crucial issue, especially if the users have never worked as a group before. For example, this might involve establishing responsibilities for making inputs and schedules for producing outputs. Decisions may also be made about the sorts of documents that should be submitted to the system, or about data traffic and classification (Hettinga, 2002).

We would emphasize that we do not see these steps as necessarily taking place in this order. However, the breakdown of the process into these steps is a logical one, and helps to understand the learning processes.

RESEARCH METHODOLOGY

Having conceptualized IT implementation as experiential learning, we should next justify the choice for interpretive research methods in this study. Firstly, an interpretive view corresponds fully with the theory of learning that we applied to IT implementation. It represents the idea that learners—the users of the technology—do not accept “the truth” about the IT as offered by the managers and project leaders, but actively construct their own views of it through their own experiences. Secondly, when talking about users’ interpretations of a technology, these are seen as reflecting the nonstatic and repeated developments in their communications towards understanding and using a system.

A case study was conducted in order to provide examples to support (or contradict) the theoretical discussion on the implementation of IT as learning, and to clarify the contents of the learning processes. We did not set strict criteria for the company to be involved in this project, but it was important that it had a recently introduced information technology. In terms of the unit of analysis, the investigation focused on the end users within an IT project. The case study techniques, that is, the research instruments, strived to build a platform for interpretive analysis and included document analysis, interviews, and participatory observations.

Collecting Empirical Data

We carried out a six-month case study in a Dutch university that was in the process of introducing SAP_HR. Data was collected using qualitative methods: semi-structured interviews, observations, and document analysis. The 24 interviews each lasted from one to 1 ½ hours. Five university HRM units were investigated in the research: the Service Centre HRM (SC_HRM), the Faculty of Social Sciences HRM (SS_HRM), the Faculty of Arts HRM (A_HRM), the Faculty of Geographical Sciences HRM (GS_HRM), and the Veterinary Laboratory HRM (AL_HRM).

Representatives of three groups of SAP_HR users were interviewed:

• Five employees involved in steering the project in the university, referred to here as project team members. They provided support for end users, carried out help-desk duties, maintained the functional and technical administration of the system, and analyzed on-going use of the system.

• Four leaders from the various faculties’ HRM departments who were responsible for the personnel policy and administration. These were not active end users of the system themselves, but SAP implementation did bring changes to their departments.

• Fifteen end users: four salary administrators from the central Salary Department and eleven HR specialists from five HRM
departments, including five key users who were seen as advanced users of the SAP_HR system. In this way we interviewed HRM specialists whose daily tasks had to be performed through the system.

The questions put during the interviews were in line with the operationalization shown in Table 1.

Transcripts of the interviews were checked and corrected by the respondents involved. Additional information was obtained during informal conversations and through participating in key-user meetings. We also studied relevant documents, namely, the Development Plan of the University “Perspectief2010,” the Project Plan and the Fit/Gap Analysis for the SAP_HR implementation, the plan for the SAP_HR pilot implementation, reports and notes from key-user meetings (04.02.2003, 18.02.2003, 04.03.2003), a special issue of the university newspaper (N26, 2001/02), plus the main manual and 36 sub-manuals covering the use of SAP_HR for the University.

Data Analysis

The collected data were analyzed using discourse analysis. Our primary concern was the social context of the technology use and the discourse that supported it. We distinguished four steps in our interpretation of the data (Van Dijk, 1997; Oevermann, 1996; Titscher, Meyer, Wodak & Vetter, 2000).

The first step involved gaining an overall impression of the views presented in the interviews and linking this to the context that was derived from the documents and observations. Knowledge of the context was crucial in order to understand and feel the implementation of SAP_HR. The second step aimed to describe the learning processes on the basis of text units from the interview transcripts. To achieve this, we had to distinguish and codify the text units from all the interview transcripts on the basis of our operationalization scheme. We used a qualitative approach along the lines indicated in the operationalization scheme (“strong – weak,” “high – low,” etc.).

Table 1. Operationalization of IT implementation as a learning process

<table>
<thead>
<tr>
<th>Learning Dimensions / Range</th>
<th>Components</th>
</tr>
</thead>
</table>
| 1. Learning as acting: task-related operations undertaken with the system by users/ Active - Passive | • operating with basic modules in performing everyday tasks
• searching for new techniques in the system |
| 2. Learning as reflecting: communicating upon the extent to which the system supports task performance/ Strong - Weak | • discussing difficulties in using the system
• comparing with other software experiences
• declaring individual problems in using the system |
| 3. Learning as conceptualizing: the level of common meaning of the system regarding its role and functionality/ High - Low | • clarity about the purpose of the system
• understanding the operation of the modules in the system
• attitudes towards system functionality |
| 4. Learning as planning: activities that aim to reach agreements on on-going use of the system / Strong - Weak | • arranging additional learning activities to improve use of the system
• developing policies
• evaluating intermediate results |
Next, the third step involved identifying the significance and linguistic representation of each text unit. We wanted to reveal a range of semantic features such as vagueness of opinions, doubts, clarity, and hidden meanings, but also the factual representation of the text units for each component. Finally, the fourth step involved refining our conclusions.

**FINDINGS: CONTEXT OF SAP_HR IMPLEMENTATION**

**Background of the SAP_HR Project**

The university had a long history as a knowledge center in the fields of scientific research and higher education. Today, it is one of the largest institutions in the Netherlands, with more than 23,000 students, and 7,000 employees. It has 18 schools and 15 support and administrative units.

By 1994, most of the faculties in the university were using a COMI-P personnel information system that was becoming outdated. The supplier no longer guaranteed updates or further development of COMI-P, and therefore there was felt a need to look for a new personnel IT system. In 1998, the directors of the faculties and other services determined the functional demands for the new system, which had to meet the following basic requirements:

- To have the ability to be integrated with the existing financial and salary administration software packages;
- to have clear and well-designed functionalities in the standard version; and
- to be easily adopted and implemented in the university environment.

On the basis of these specifications, the University Board made the decision in November 2000 to purchase the SAP_HR personnel management system. At that point, the university was already using SAP_Financial, the financial module from SAP®. In choosing SAP_HR, the organization hoped for a painless implementation trajectory based on their existing experience with SAP_Financial, and expected to achieve an easy match between their HR and financial administrations.

**The SAP_HR Introduction in the University**

In the annual university report on IT (2002), under the heading “Administration and IT” one can read the following statement:

*The implementation of the personnel module within the SAP system reveals more outset problems than was expected, especially in management information reports and matching with the salary administration system.*

Problems in matching with the salary administration system” in fact meant delays, mistakes, and other difficulties in the payment of salaries. The university’s weekly newspaper published a small article that highlighted such problems (dated 14 March 2002, # 26):

*Tens of employees have got less salary than they should have... Especially those who had any changes in their contracts since 1 January 2002 and those who had multiple short-term contracts. On the other hand, those employees whose contracts expired on the 1st of January 2002 continued to be paid... According to the project team, the origins of the problems are too complex.*

A historical account of the SAP_HR implementation project is given below.

In December 2000, the university started searching for a consultancy firm to help with the
Successes and Failures of SAP Implementation

In April 2001, a consultancy firm was chosen on the basis of its experience with both SAP® and IPA technologies.

The period April through December 2001 was expected to see a “fast implementation” of SAP_HR. The steering group looked for discrepancies between SAP_HR and the existing systems. Six project groups worked on different aspects of the implementation: realization, salary/IPA, acceptance, technique, conversion, and training.

One unforeseen event which interfered with the project was that the expert from the consultancy firm who specialized in developing an interface with the IPA external salary system left the project. In October 2001, pilots took place in four units: the faculties of biology, pharmacy, and chemistry, and the P&O Service Centre. During November and December 2001, all the future users of the system took training courses. Preparation was carried out on technical issues such as conversion and transportation.

On January 1st 2002, SAP_HR was introduced in 12 faculties and in all the support and administrative services in the university. Two faculties refused SAP_HR, and kept their old personnel systems. In the opinions of the users, the introduction date for the new system was not ideal since it coincided with the introduction of a new Collective Agreement for the Dutch Universities (which had to be processed through personnel administration). Further, some units were restructured requiring further new paperwork.

Documents show that an evaluation of the implementation was already scheduled for March through April 2002, that is, three to four months after the SAP_HR introduction. However, shortly after the introduction, and continuing through to the summer of 2002, unexpected difficulties arose related to inputs and outputs to SAP_HR, and with sending data to the external IPA salary system.

The extent of the drama during the first 7 to 8 months was expressed in various ways. For example, we heard of about 3,000 mistakes being registered in the database with only a third being resolved, 450 e-mails in 6 months from the users reporting problems, 75 “crucial” problems that had to be resolved, 10-20 technical changes/improvements per day, and finally about 300-400 university employees who experienced problems in getting their salaries.

The first months were really terrible. We made inputs in accordance with our experiences and the knowledge we got from the course, but most of the time there were mistakes, and IPA rejected the data. As a result, employees did not get their salary. Sometimes this went on for some months. Mistakes could be very simple and unexpected, but they took a long time to identify. (Roy, A_P&O)

Now I am sure—if they want to do something like this again in the same way—I am leaving. I really mean that! It was just one big disaster from the beginning. People did not get any income for three months. It was terrible and unclear who was responsible for what. Many HRM specialists became sick…. (Erik, SS_P&O)

We did not observe a “happy ending” to the SAP_HR implementation phase during our 6 months of involvement. However, at least the number of employees experiencing problems with their salaries had decreased from 300-400 in spring 2002 to 60-100 by March 2003.

Tasks and Responsibilities of the SAP_HR Users

After SAP_HR was introduced on January 1st 2002, the various HRM specialists were compelled to collaborate with each other. In total, there were about 50 users who worked in six different university units: four HRM departments from the faculties, the Service Centre’s HRM, and the central Salary Department. They were located in
various places: two in the administrative building on the university campus, three in other buildings on the same campus, and one in the historical center of the city. All the users had their own managers, either in their own HRM units or in the salary department as appropriate.

In terms of the SAP_HR implementation, each unit had two types of users: “regular” users and “key” users. Key users were responsible for correspondence with the project team, helping “regular” users, searching for new possibilities in the system, and attending special meetings.

At the time of our research, the personnel administrators were busy processing changes in the personnel files of the university employees. These files were either paper-based or based on SAP_HR. We found about 40 tasks performed using SAP_HR that could be grouped into ten categories: (1) appointment of an employee (subtasks concerning the appointment of a new employee or an external worker, stagier, and those with nil-contracts); (2) modification of basic information, payment information, working time registration and other data; (3) relocation processing; (4) promotion; (5) work time registration; (6) administration of leave (sabbatical, sick, parental, and pregnancy); (7) processing an optional model for employment conditions which is only in part executed through SAP_HR; (8) administration of declarations; (9) vacations; and (10) producing HR statistical reports and information management reports (sick leave reports, HR financial reports, etc.).

We observed some variations among the five units in terms of:

- The number of employees working in the HRM units: SC_HRM (11 employees), SS_HRM (9 employees), A_HRM (9 employees), GS_HRM (5 employees), AL_HRM (1 employee);
- In some of the HRM units, all the employees would perform all the HR administration-related tasks, while in others some were only responsible for communicating with the employees of the faculty while others performed key user tasks with SAP_HR; and
- The functions of key users were assigned differently. For example, in SC_HRM, all employees could represent the unit as its key user, while in others there were strict divisions.

Each faculty had its own special characteristics that influenced the use of SAP_HR in its HR administration. For example, the Service Centre was a special structure within the university that provided HRM services to more than 400 employees including those in three smaller faculties and more than 20 administrative and support services such as the university library, museum, and communication department. Another example is from the “GS” faculty where there were many short-term workers who did not work fixed hours and did not receive a regular salary. Student assistants generally fell within this category (flex workers) in this faculty as they were usually appointed to execute specific tasks within projects. Of about 320 employees, 1/4 were classed as flex-workers.

These examples indicate how the task needs and divisions differed per unit based upon the “idiosyncrasies” of the faculties. As a result, the ways used to process HR information also varied, and these differences needed to be acknowledged in SAP_HR.

Changes Brought by SAP_HR for the Users

The interviews with the SAP_HR users showed that an apparently straightforward technical inten-
tion had brought with it many social changes. We have clustered these into three groups:

1. HRM administrators had increased responsibilities for the transactions they completed; as one of them noted:

   *With SAP we gained extra control, but also more responsibilities. We have to be very careful with all inputs. Earlier everything was on paper, but now we have to concentrate more intensively in order to avoid faults.* (Roy, A_HRM)

2. Task interdependence had changed radically. Instead of being concerned only with internal paperwork within the faculties, all the inputs made by personnel administrators now became interdependent with the inputs of salary administrators, and over time with the IPA Salary Information System that was located outside the organization.

3. Unlike the old situation, online working with personnel and salary documents demanded standardization in the operationalization of the personnel and salary tasks and processes by the entire group of users. Whereas, as we have already discussed, the HRM units all had their own traditions and rules in the pre-SAP situation, the new circumstances required clear definitions of all the terms and processes used. This reinforced interdependency between all the units.

**LEARNING IN SAP_HR IMPLEMENTATION**

We were able to distinguish two distinct periods in the implementation of SAP_HR: the first 6 to 8 months which were described as a “disaster” (time-1), and later when the situation had improved (time-2).

**Learning as Acting**

When they began to work with the system, the users strived to handle basic tasks such as inputting personnel data, sick leave administration, time registration, and applying the right types of contracts through SAP_HR. They expressed the view that they were initially afraid to work with the system because they could not predict whether many of the transactions would be correct or not. In such cases, they preferred to contact a salary specialist or a key user and ask them to execute the task.

The users had to operate the system because it was necessary in order to perform their primary tasks. However, the intensity of use did differ from unit to unit. For example, based upon the interviewees’ estimates, the Salary Department processed about 250 transactions per week and the A_HRM unit about 250 transactions per month, whereas the user in AL_HRM worked for no more than two hours per week with SAP_HR.

Initially, the users sensed a lack of time or motivation to search for new possibilities within SAP_HR but, after 6 months, the situation changed: they could now work with the system without asking for help every time, and they could begin to search for new possibilities in SAP_HR. For example, there was a special application called “Query” through which a user could generate a range of HR reports. The interviewees emphasized that they found it interesting to combine HR and financial data. All the key users had test versions of SAP_HR, with which they could experiment and search out new possibilities and produce reports. However, it was commented that these versions did not indicate any link to IPA, something that was essential for much of the work.

**Learning as Reflecting**

The interviewees expressed the view that, initially, there were no fruitful communications across the
entire group of users. Opinions were expressed that no one wanted to admit their own mistakes and so always blamed others, for example:

*Sometimes it was not only technical difficulties that caused the problematic situation. Correct and timely communication is very important. Even within those groups closely related to the salary administration groups we cannot always find consensus: when anything goes wrong, everybody is sure that they did their own job well, and the problem must be elsewhere. Such communications do not help to improve the situation, and we might face similar difficulties in the future.* (Daniel, SAP technical administrator)

Lack of time was considered as the main reason for the lack of cross-communication, for example:

*We did not communicate with the HRM units about the use of the system. We did not even think about that—there were so many mistakes that had to be corrected; it was easier to do this ourselves instead of talking with the HRMs. It was terrible that we had to correct all the inputs.* (Karen, Salary Administration).

However, within the units, there were active discussions about the problems with the SAP_HR administration system. In the A_HRM unit, meetings took place every two weeks, and in SS_HRM every week. The personnel administrator from GS_HRM described the situation as follows:

*We [the Personnel Department] worked together very well. We discussed difficulties and helped each other with this system. We made reports about mistakes ourselves, and the key-user took these to the regular meetings. In our faculty, we are lucky to have such a strong team. With all these SAP problems we became even closer to each other.* (Tom, GS_HRM)

Gradually, after some months of working with SAP_HR, the users from the various units became more open in the discussions. They expressed enthusiasm for communicating across the entire group in this later phase of SAP_HR use:

*Also we communicate with other HR managers to ask questions and share difficulties. In this way, people from the Service Centre helped us a lot at the beginning. We also found it useful to discuss SAP with the HRMs from the Social Sciences faculty.* (Roy, A_HRM)

Key-user meetings became an important event for information exchange. The key users took the latest news to and from the meetings; and users started sending e-mails across the group with their questions. The meetings of key users became a strong group device; even non-key users attended in order to participate in the communication process. During the interviews, all the respondents acknowledged the importance of these meetings:

*I like communicating with other users. During the key-user meetings we raise a range of questions and exchange our ideas. It is very helpful. Actually I am not a key-user, but I like to attend those meetings (together with the ‘real’ key-user from our unit) to gather all the news and to communicate with others. There, I always meet the Salary Administration people and talk with them. I also visit them after each meeting—to chat face-to-face—otherwise we would communicate only by telephone.* (Marijke, A_HRM)

### Learning as Conceptualizing

All the interviewees were well-informed and understood the goals behind introducing SAP_HR. They gave two main objectives as being behind SAP_HR’s introduction: replacing an outdated system, and matching the SAP modules already
Successes and Failures of SAP Implementation

being used in the university. Those involved in the working project groups were all of the opinion that, by January 2006, the IPA system would be replaced by SAP_Payroll, and thought that SAP functionality would then become even more valuable.

The comment below from one of the interviewees shows the clarity of his awareness of the goals behind SAP_HR:

*The system was introduced in January 2002 because of two reasons. Firstly, there was already SAP_Financial, and the financial department had worked with SAP for some time. I think the management desired to have the university’s ICT from a single supplier. Another reason for the introduction of SAP_HR was that the contract with COMI-P had expired by January 1st 2002. (Roy, A_HRM).*

None of the users however expressed a need for a new system and none of them felt an urgent need to replace the “old” technology. On the contrary, they stressed that the previous technology was reliable enough, simple, and worked correctly. An interviewee from the AL_HRM department gave several reasons why she did not need SAP_HR:

*I think SAP_HR is a good system. You can do many things with it, but I don’t need many things. For example, we have our own system for sick leave administration. The same applies to time registration - there is our internal ATREA system. This contains various special items such as overtime, working during the weekends or holidays, and evening work. It has existed for ten years already. Maybe it can be incorporated into SAP, I don’t know. Therefore, I don’t use the sick leave administration and time registration components in SAP_HR. I don’t use the ‘arrangements’ application either. They do this task in the performance appraisal files and keep them on paper. With SAP, this would be extra work for me. Other examples of useless applications are the ‘previous employer’ field, and the ‘children’ and ‘subscriptions’ fields - I don’t need them. (Monique, AL_HRM)*

During the first 6 months, the users felt that they did not really understand how to operate SAP_HR. All the 24 interviewees commented that they lacked an understanding of the logic of the system. For example, a salary administrator said:

*It was terrible that we had to correct inputs, and that we did not have enough knowledge about the system and how to work with it. We did not even have an image of a good input, and how a correct input should look. It was very confusing for us because one month an input ‘A’ was good and accepted by the IPA system, but the next month the same input ‘A’ was certified as bad and rejected by the same IPA. It was not clear what went on behind the screen. (Karen, Salary Administration)*

The main complaints were about the lack of understanding of what was “behind the screen.” It was not difficult to click the buttons, but they needed to foresee the outputs of the transactions and the connection with IPA which, at the beginning, seemed to be a big black box.

Based on many of the opinions expressed, the situation at the beginning could be characterized as one of high uncertainty. Most problems and their understanding would only come from experience; they could not be predicted in advance:

*The situation at the beginning could, in general, be characterized as one of high uncertainty – COMI-P was very rapidly replaced with SAP_HR. We got a new system, but we did not know sufficiently what to do. The biggest problem, and the highest priority, was to keep to the transaction deadlines. (Sandra, SS_HRM)*
In fact, none of the project leaders realized that we — the HRM specialists — did not know about IPA. We had never worked with it. The end-users in their day-to-day work only see SAP screens. We were often confused because sometimes SAP_HR would allow us to input a number (as a code), but then it was then forbidden by IPA, etc. (Lucie, GS_HRM).

In assessing the users’ attitudes towards the functionality of the system, we note that most opinions were negative. Criticisms concerned both technical and contextual aspects of SAP_HR. We have produced the following summary of criticisms that arose during the interviews:

- making mistakes was “blind”: a user could not understand why an input was wrong;
- some problems were too difficult to solve;
- employee classification in the system was too complex;
- searching for new possibilities was limited as the system was very standardized;
- producing historical overviews was impossible;
- the codes in SAP_HR were different from the codes in IPA, and therefore there was a need to memorize them;
- useless functionalities (such as inputting educational data on the employees and data on their children which was not processed in calculating salary); and
- some issues that are common in a university environment were not incorporated in SAP_HR (conference leave, sabbatical leave, and CAO à la carte).

One of the personnel administrators described her attitudes towards SAP_HR as follows:

In April 2002 I started to hate both the system and working with it. I had a feeling that everything I did went wrong, and that it was all about salaries and bonuses. (Monique, AL_HRM)

At the same time, the functional and technical administrators of the system were of the opinion that SAP_HR was very logical, technically reliable, and easy to use.

**Learning as Planning**

Arranging activities to improve system use became an observable occurrence after several months of practical experience. In the beginning, activities, if any, were initiated by the project team and not by the users. However, since spring 2002, as the interviewees themselves noted, the users have tried to initiate actions in addition to the key-user meetings. Thus, “informal discussions over a cup of tea” were arranged between the SS_HRM unit and the Salary Department. Discussions about certain transactions were also initiated outside of official meetings (e.g., the development of a report for the Executive Board).

We also discovered differences in the policies developed by the users in the various units:

- control over transactions was organized in various ways, from triple control involving the head of the HRM unit, to double checking by the same user in AL_HRM;
- in GS_HRM, there was an agreement that the key user would decide whether to inform regular users about e-mails from the project team or not, in order not to “overload” them;
- each HRM unit had its own time schedule within its faculty for making changes to personnel files, and they agreed a schedule with the Salary Department for providing them with the data that would guarantee timely salary payments; and
- in January/February 2003 (a year after the system was introduced), the Salary Department introduced “report forms” for those HRMs who had questions/problems in order to initiate discussions rather than just correcting the mistakes themselves.
All the interviewees noted that there were no evaluation rounds throughout the project.

**DISCUSSION**

**Success of SAP_HR Implementation and the Role of Learning Processes**

When we started the case study we knew that the SAP_HR implementation had developed many problems at the university. The users had struggled with problems associated with SAP_HR for 6 to 8 months before working with the technology became easier although still not fully enjoyable. After this initial period, the employees became more willing to cooperate with each other in order to develop their work with the technology. How did this develop?

After 8 months of preparation, the technology was introduced to the users on January 1st 2002. Although the users received instructions and participated in workshops about setting up and converting to SAP_HR prior to its introduction, they found that they were not ready to operate the system but lacked the option of rejecting it.

The introduction of SAP_HR was initiated and promoted by the top management in the university, and this choice of technology was never fully supported by the future users. SAP_HR use became obligatory with the objective of replacing what was seen in some quarters as the outdated COMI-P system and so standardizing HRM and salary administration in the organization.

What did the users experience and feel once SAP_HR had been introduced to them? Personnel administrators saw significant changes in their daily tasks: greater responsibilities for making online inputs, and more control over those inputs, the need to become interdependent with the salary administrators, and a need to collaborate with other personnel administrators whom they did not previously know. Salary administrators also gained new tasks: to control the inputs from the HRM departments, to collaborate with them, and to learn how to operate SAP_HR. One additional issue that complicated working with SAP_HR was that the interface with the external IPA salary system often obscured the SAP_HR inputs.

Stress, greater responsibilities, and uncertainty in making inputs, all brought about by SAP_HR, stimulated negative interpretive schemes about the technology among the users. They were unwilling to invest a lot of effort and were disappointed by the technology. Right from the beginning, the users perceived the system as not worth learning about, and in fact, worse than the previous technology. The negative opinions about SAP_HR strengthened daily as the users collected and accumulated disappointments, including seemingly small details and misunderstandings with the project team.

Neither did we see strong user participation in the project. Only the key users actively took part in the preparation and conversion of the system. A pilot for introducing the system took place in four units in the university. Although the users did not agree with the official, positive evaluation of this, their opinions were largely ignored.

Although a negative opinion about SAP_HR grew within the user group, use of the provided technology remained obligatory for the staff. Slowly, after 6 to 8 months, the interpretations of SAP_HR began to move in a positive direction. The employees started to find ways of avoiding or overcoming the major problems they had found with SAP_HR.

The question is how does a group learning perspective help us to explain the developments in the SAP_HR implementation? We saw how group learning among the SAP_HR users developed from an initially low level to moderate after 8 months of using the system. This development was slow and this created difficulties for the users. The qualitative analysis of the transcribed interviews and documents has enabled us to distinguish and grade all the learning processes twice, at the beginning of the SAP_HR implementation and after 6 to 8 months (see Table 2).
The four learning processes all progressed in a positive direction over time although mutual adjustment processes progressed only slowly.

In terms of acting, the users progressed from being afraid to click the buttons at the beginning to attempting to generate a range of HR reports. The key users remained the most active, but the rest of the group also became more active. Everybody found routine administrative tasks easier to execute over time. Reflection among the targeted employees also developed progressively. Initially, discussions about SAP_HR implementation took place only at the microlevel, that is, within the units, and there were no fruitful interunit communications, not even between Salary Administration and the HRM departments. However, later, this developed with e-mail, telephone, and other informal ways of corresponding and discussing SAP_HR taking place across the units. The key-user meetings became especially popular.

The general understanding of the purpose of the system coincided with the reality. All the users were informed about the goals of SAP_HR and could express them correctly. However, they perceived it as useless, and they did not feel there had been any immediate need for a technological change in their tasks. We did not find users expressing needs in terms of SAP_HR, even during the later stages of working with it. Users’ attitudes towards the functionality did not improve during the observed period and remained negative. Only key users appreciated the possibilities of generating reports. We did not discover any activities arranged by the ordinary users in order to improve their work with SAP_HR, in fact there were only two informal meetings and these were initiated by the SS_HRM department. Most policies were developed at the microlevel in the units (such as rules to control the inputs or processing CAO à la carte). There were no evaluation rounds established to assess the system, the project, or cooperation.

We found that, at the beginning of the project, the learning processes had mainly occurred at the microlevel within the units. Further, the interaction processes across the entire group of users were at a very low level during the first months following implementation. We attribute this to the initial lack of shared structural and nonstructural group features.

CONCLUSION

Our case study concerned the implementation of a personnel administration system in a university environment. A six-month investigation
Period allowed insights to be gained into the processes involved in adopting the system by the user group.

Learning among the SAP HR users emerged as soon as the system was introduced. We observed the slow but steady development of all the learning processes over time. Initially, the extent of the learning was not sufficient to handle the system, with interactions mostly taking place at the level of the units. The lack of communication and the misunderstandings across the entire group of users led to mistakes in working with the system when it was first introduced. The system triggered learning by calling for a redirection of all processes towards a new and larger community. After 8 months, learning moved towards stronger cross-unit cooperation and the exchange of users’ experiences.

We began the theoretical discussion with an understanding of IT implementation as a user-centered process in which employees together develop interpretive schemes about a newly introduced technology. In this study, we based the concept of learning on the model of experiential learning by Kolb (1984) in which learning is considered as: (1) a process rather than simply outcomes; (2) a problem-solving process that is always practice-oriented; and (3) a mechanism for everyday activities, occurring both consciously and unconsciously. It is argued that learning within a group of users is more than simply the summation of individual learning processes. The character of group processes becomes more complex as they acquire a social context.

In answering the central research question – What is the role of experiential learning in IT implementation? – we would stress four issues: (a) learning is a process-based activity; (b) it rests on the interaction processes between members of a user group; (c) these processes begin when a new technology is introduced.; and (d) these processes lead to changes in knowledge about the IT and in users’ behaviors (ways of operating the system).

The main role/function of learning processes in IT implementation is that they become a “hidden” mechanism in the IT implementation. Several theoretical components can be added here to articulate the strategic potential of experiential learning in IT implementation. First, there are two learning processes – acting and reflecting – that concern individual behavior. Observations have shown that these occur immediately after a technology is introduced to the targeted employees. Once users are “thrown to chaos” (Weick, Sutcliffe & Obstfeld, 2005), they have to (or choose to) experience it, and their next implicit step will be to judge it. The other two processes—conceptualizing and planning—bridge individual and group learning. These two steps concern “talking a situation into existence” (Taylor & Emery, 2000) and articulating and integrating labels.

We observed that verbally expressing concerns developed more extensively when groups of users established such characteristics as trust, knowing each other, and openness in risk-taking conversations (known as psychological safety) (Edmondson, 1999). These characteristics were seen to develop during the IT project.

Findings elsewhere have shown that group learning processes do not follow a linear sequence, but develop as cycles. Within a group, discussions occur both before and after actions and so it is difficult to distinguish a “point of departure” in an experiential group learning cycle.

Finally, we argue that learning can be considered as a “hidden” mechanism for speeding up or slowing implementation. As Weick et al. (2005) remarked, it is a micromechanism that can bring macro-changes. If a group of users appreciate the technological help provided for their tasks, share positive attitudes, help each other, and attribute growth in performance to the system, they will learn the relevant issues concerning a technology. In so doing, the system will be discovered, “studied,” and better understood, and through this the technology will become more relevant for the job tasks and easier to work with. This can lead to a
Successes and Failures of SAP Implementation

better and quicker acceptance by the users. The opposite scenario was initially observed in our case study when the users only complained about the system, perceived it negatively, and convinced each other of its uselessness. Even small details that would be ignored in other situations received the attention of the group. In such a scenario, employees learn of issues that reinforce earlier negative attitudes. Their views became increasingly negative about the relevance of the system to their tasks, and they saw the technology as too complex to operate. Thus, the technology became “even less relevant” for their work in the opinions of the users, and they learned of issues that discouraged them from accepting the system.

Further research could greatly contribute to the understanding of the origins and differences in learning by taking into account differences in work environments. Insights could be gained by exploring IT implementation in various types of work environments (such as process-based, product-based, logistics-based, and administrative work). Determining whether there is a link between the type of work environment or the type of organization, group learning in IT implementation would add to the current findings.

Practical Implications

Our findings suggest that the main thrust of managerial support during the implementation of information technologies should be in promoting interaction processes geared towards adopting the system. We observed a number of good practices in the case study that did stimulate constructive learning. These were:

- having a help desk or front/back office services on system functionality available for the users at any time;
- creating and distributing a list of experts on the system’s functionality within the group (generally these were the advanced users among the targeted employees) whose experience could be very helpful to others;
- introducing an e-mailing list including all the users (setting up a hot-line chat room would be an alternative);
- scheduling informal meetings (such as during coffee breaks) for the group of users;
- agreeing how to involve new employees in using the system (what to explain to them, who is responsible, etc.);
- distributing special notebooks amongst the users for ideas, proposals, and complaints; and
- accepting proposals that come from the users and reacting to them (negotiating).

This list is far from exhaustive since it includes only those practices that we observed in the implementation of the SAP_HR system at the university. However, we did observe how all the above practices advanced learning in the “right” direction.

REFERENCES


*This work was previously published in the International Journal of Technology and Human Interaction, edited by B. C. Stahl, Volume 3, Issue 4, pp. 33-51, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 6.16
The Human Face of E–Business: Engendering Consumer Initial Trust Through the Use of Images of Sales Personnel on E–Commerce Web Sites

Khalid Aldiri
University of Bradford, UK

Dave Hobbs
University of Bradford, UK

Rami Qahwaji
University of Bradford, UK

ABSTRACT
Business-to-consumer (B2C) e-commerce suffers from consumers’ lack of trust. This may be partly attributable to the lack of face-to-face interpersonal exchanges that provide trust behavior in conventional commerce. It was proposed that initial trust may be built by simulating face-to-face interaction. To test this, an extensive laboratory-based experiment was conducted to assess the initial trust in consumers using four online vendors’ Web sites with a variety of still and video images of sales personnel, both Western and Saudi Arabian. Initial trust was found to be enhanced for Web sites employing photographs and video clips compared to control Web sites lacking such images; also the effect of culture was stronger in the Saudi Arabian setting when using Saudi photos rather than Western photos.

INTRODUCTION
The beginning of the 21st century brought rapid development to the field of e-commerce and many enterprises in Western developed countries found success in this area. According to emarketer.com, total online retail sales for 2005 were $144,613 million. In 2001 Internet sales to households from the UK non-financial sector stood at £4 billion;
by 2004 these had increased to over £18 billion (Wallis, 2006). The growth in Internet sales between 2003 and 2004 was over 67%; this clearly shows brisk expansion in the value of Internet sales (Wallis, 2006). In the Asia-Pacific region, sales also grew rapidly, from about $200 billion in 2003 to about $300 billion by 2004 (United Nations Conference on Trade and Development [UNCTAD], 2003). In contrast, the situation in the Arab world is different since the estimated figure for B2C in Africa and the Middle East cannot be compared with those in the US, the EU, or China (UNCTAD, 2004).

Trust in online shopping still represents a significant barrier for Internet users and is crippling the e-commerce environment (Zhang & Zhang, 2005). Trust is a critical component for any business transaction, and is particularly essential in the e-commerce environment because of the paucity of rules and customs in regulating e-commerce and due to the fact that, typically, online services and products are not immediately verifiable. Moreover, online transactions are more impersonal, anonymous, and automated and lack the assurance provided in traditional settings through formal proceedings and receipts. B2C online vendors are particularly challenged when trying to build and attain consumer trust. Indeed, in contrast to face-to-face commerce and to other applications of trust, which encompass a wide range of emotions involving various types of social interaction with humans, there are typically no interpersonal interactions in e-commerce, neither direct nor implied. Such interactions, or even cues relating to them, are notably missing from e-commerce Web sites. Online shopping may be viewed as lacking human warmth and sociability. The social aspect of shopping has been shown to be a major contributor towards positive emotions (Jones, 1999) that have been linked to several important outcomes, such as increased unplanned purchasing, increased spending, and increased time spent in the store (Jones, 1999).

Online vendors face a significant challenge in making their Web sites socially rich (Kumar & Benbasat, 2002).

The aspect of trust has been examined over many years, however, most of the research on consumer trust focuses on consumers in English-speaking countries and newly industrialized countries. Trust in e-commerce research has been conducted almost exclusively in the US or Western Europe. Considering the cultural, economic, social, and political differences between the developed and less developed countries, concerns have been raised in terms of whether research based in the US or Western Europe can be generalized to other countries (Bagozzi, Wong, Abe, & Bergami, 2000). The global nature of the Internet raises questions about the effects of trust across cultures as well. Although trust may be formed in a variety of ways, it depends on the cultural factors (e.g., societal norms, values, etc.) that guide people’s behaviors and beliefs. Despite the differences national culture can cause in e-commerce behavior (Kacen & Lee, 2002), and despite e-commerce becoming global, research on trust in e-commerce has to emphasize the possible effects of national culture. For that, this study attempts to carry out a pilot case investigation of our research objectives in Saudi Arabia; a developing country with the highest growth in Internet users in the Middle East and Africa (www.mcit.gov.sa).

Research Objectives and Organization of the Article

This study researched the effect of adding media cues (video clips, photographs) of a presumed sales person’s face to the home page of an e-commerce vendor, and considered the effect of social presence cues, also known as interpersonal cues, that are implicit in the media cue on users’ initial trust in the vendor (Research Objective 1). The majority of studies of online customer trust focus
on general trust as it develops between customers and online vendors over time and after repeated experiences. While we recognize the importance of the evolving nature of trust, our study focuses on initial trust beliefs that develop after a customer first interacts with the company’s Web site. The difference between the two concepts may be subtle but, as we discuss later, is significant, particularly in the online environment. The study further investigated the effect of culture, as represented by a facial photograph of a Western and Saudi man added to the home page of an e-commerce vendor, on users’ trust in that vendor (Research Objective 2). This experiment was concerned with investigating trust as a dependent variable and other variables as independent variables, as well as how the depth of a user’s exploration of a site (exploration depth) influences the effect of a media cue on user trust (Research Objective 3). On a methodological level, the experiment aimed to overcome part of the limitations of conventionally used trust questionnaires free from financial risk by introducing trust questionnaires that were elicited under conditions of financial risk (Research Objective 4). This article is organized as follows: the second section examines previous work related to e-commerce and trust and reviews the research hypotheses. The third section outlines the methodology of the study, while the data analysis is presented in the fourth section. The results are discussed in the fifth section and in the final section conclusions are drawn.

THEORETICAL BACKGROUND AND HYPOTHESES

Trust

Trust has widely conflicting conceptual definitions; the literature on trust in general, and on trust in e-commerce in particular, is in a state of confusion (McKnight & Chervany, 2001). The confusion in trust terminology is expressed in terms of two kinds of problems. First, similar concepts are given different names and second, the same terms are used for different concepts (Krauter, Kaluscha, & Fladnitzer, 2006). There is no universally accepted scholarly definition of trust (Rousseau, Sitkin, Butt, & Camerer, 1998).

Summarizing the different definitions of trust across various research disciplines, it can be concluded that all trust definitions address one or more of the following perspectives: (1) context characteristics, (2) trustor properties, and (3) characteristics of the trusted object (Krauter et al., 2006). Many definitions also address the interaction or relationships between two or all three of these perspectives. The key concepts of most of the trust definitions are risk (Mayer, Davis, & Schoorman, 1995), vulnerability (Mayer et al., 1995), expectation (Baier, 1986), and confidence (Lewicki & Bunker, 1995). These factors are gathered by the following trust definition, which represent the most common definition of online trust (Mayer et al., 1995):

Trust is the willingness to depend upon another party and be vulnerable to the actions of this other party based on the assumption that the trusted party will deliver without taking advantage of the situation. (p. 711)

As it is very abstract, this definition can be seen as the most common in the human computer interaction (HCI) trust research (Corritore, Kracher, & Wiedenbeck, 2003). It defines trust as an internal state or attitude which entails cognitive and affective aspects (Corritore et al., 2003). Trust, as an internal state, is different from trusting action (e.g., buying online, entering credit card details, relying on advice) or cooperation (e.g., in a team setting), which are observable behaviors (Corritore et al., 2003). Also trust is not the same as trustworthiness; trust is the act of a trustor while perceived trustworthiness is a characteristic of someone or something that is the object of trust (Corritore et
In the e-commerce context, trust has been defined rather narrowly (e.g., Bhattacharjee, 2002, defined trust as trusting beliefs) or more broadly (e.g., McKnight, Choudhury, & Kacmar, 2002, who defined trust as trusting beliefs and trusting intentions). Trusting beliefs (i.e., perceptions of Web vendor attributes), and trusting intentions (i.e., intentions to engage in trust-related behaviors with a Web vendor) are based on the theory of reasoned action (TRA) which was introduced by Ajzen and Fishbein (1980) to analyze the psychological processes that reflect observed relationships among beliefs, attitudes, intentions, and behaviors. The theory asserts that intention to perform behavior is determined by the individual's attitude toward the behavior, and a person's attitude is affected by his/her beliefs. TRA has been widely accepted and applied to a broad range of disciplines and contexts. Existing empirical research has revealed that trust is significantly related to attitude, and attitude positively signifies people's purchase intentions (Chow & Holden, 1997; Macintosh & Lockshin, 1997). TRA is also applied as the theoretical base in recent studies on trust formation (McKnight & Chervany, 2001; McKnight, Cummings, & Chervany, 1998), especially in the context of e-commerce (Jarvenpaa, Tractinsky, & Vitale, 2000). Since trust can be seen as a belief, confidence, sentiment, or expectation about an exchange partner's intention and/or likely behavior, we believe that it is posited to be directly related to the attitudes toward purchasing from a vendor and indirectly related to consumers' willingness to buy through purchasing attitudes. This is how we define trust in this article; this is consistent with how trust has been previously defined in the literature (Gefen, 2002).

Trustors and trustees, that is, objects of trust, can be individual people, groups, organizations, or even societies. This article focuses on the narrow-scale study of trust, where the target is an individual e-business. At this level, the focus is on the process of identifying a trustee's characteristics or his trustworthiness.

The Varying Dimensionality of Trust

Although there is no complete agreement on the meaning of trust, it should be noted that the dimensions that constitute it are context-specific (Lewis & Weigert, 1985). As mentioned previously, the literature does reveal that the concept of trust tends to coalesce around a few key recurring concepts. There must exist, for the trusting party, uncertainty about a potential or existing relationship that leads to a certain perception of risk or vulnerability. This perception of risk is generally based upon the beliefs regarding specific factors of the trustee. Yet, recent research has primarily focused on three specific factors which may parsimoniously capture the concept or the dimensions of trustworthiness—that is, ability, benevolence, and integrity (Jarvenpaa et al., 2000). Ability is the perceived skills, competencies, and characteristics that enable a party to have influence within a specific domain (Mayer et al., 1995). Benevolence is the trustor's belief that the trustee wants to do good toward the trustor. Integrity is the belief that the trustee adheres to a set of principles that the trustee finds acceptable (Mayer et al., 1995). Perceptions of these three characteristics can lead to a willingness of the trusting party to depend or rely upon the trustee in expectation of a certain beneficial outcome or that the trustee will not act opportunistically. These three specific factors can often not be observed directly, but need to be inferred from signals in the Web site interface. In HCI trust research, signals given in the interface of a system that can form the basis of perceived trustworthiness are referred to as surface cues (Fogg, 2002). Certain literatures construct trust from different viewpoints and propose two main components or factors of trust: cognitive and affective or emotional (e.g., Shneiderman, 2000). Feng, Lazar, and Preece (2004) argue that the affective component...
of trust has been to some extent ignored in HCI trust research, in particular in the domain of e-commerce, where research is largely focused on factors that contribute to cognitive trust, such as security, error absence, and trust seals. For retail transactions, cognitive factors can increase the confidence that a transaction will be successfully completed. Affective factors can increase the feelings of attraction and trust.

The distinction between cognitive and emotional trust has received some support in an empirical study on long-term work relationships (McAllister, 1995), and strong support in a face-to-face study that manipulated emotional trust (Johnson-George & Swap, 1982). However, only a few studies have examined affective-oriented factors in an e-commerce (B2C) environment.

This oversight of the affective element of trust does not reflect the increasing consideration that is given to affective processes in multiple disciplines (e.g., marketing, psychology, sociology) (Riegelsberger, Sasse, & McCarthy, 2005). For example, marketing and advertising research are the fields that have the greatest power in investigating affective elements in decision-making processes. In these disciplines it is well-known that a consumer’s purchase decision depends on rational process and on positive affective reactions (Kotler, 2002).

However, as mentioned in the research objectives, this study concentrates on initial trust. The definition of trust previously discussed also applies to initial trust. The difference lies in the temporal context of the development of trust. In this case, initial trust is the willingness to rely on a third party after the first interaction with that party. It is recognized that the development of trust is an ongoing, dynamic process that is influenced by interactions between two parties. However, we believe that initial trust beliefs can also be formed without any prior experience or interaction between the two parties. The initial trust beliefs may change with time and with repeated interactions but they may also determine the extent to which such future interactions will take place (McKnight, Choudhury, & Kacmar, 2002).

In the context of e-commerce, initial trust is very important since Web-based companies must rely on their Web sites to represent them and to show their new customers that they are reliable and trustworthy. Customers visiting a Web site for the first time may rely heavily on Web site cues to form their initial trust beliefs, but they still lack the important information that develops through interaction over time (McKnight et al., 2002). Another important reason for establishing customer trust through the first Web site visit is the possibility of low switching costs for new customers compared with switching costs for expert customers (Reichheld & Schefter, 2000). Since the Web contains vast numbers of alternatives, either in vendors or in products, it has become very easy for customers to switch to a different online store after their first visit. So it has become an essential process for online vendors to try to establish and retain a trusting relationship with first time customers. Therefore, we believe that a thorough examination of the factors that promote initial trust in new customers of online vendors is critical.

**Social Presence and Online Trust**

Applying trust-inducing features to the Web sites of online merchants is the most effective technique for enhancing online trust. Many researchers in HCI have begun to study trust in an online context. Some researchers are focusing on the cues that may affect trust or convey trustworthiness to users of Web sites. In a sense, a Web site with trust-inducing features functions as a skillful salesperson for the company (e.g., Katsikas, Lopez, & Pernul, 2005).

The trust phenomenon represents an ongoing interpersonal interaction with another party, whether a person or organization, over an extended...
period of time. Indeed, trust is typically built gradually through extensive ongoing interactions that enable individuals to create reliable expectations of what the other party may do (Luhmann, 1979). This social context is a key characteristic of trust and trust-building behaviors, since trust, in general, is built through constructive interactions with other people (Luhmann, 1979). So, the perception of a high degree of social presence, implying direct or indirect human contact in the relationship should, arguably, contribute to the building of trust. Extending this logic implies that another way in which trust in an e-commerce Web site may be built is through embedding the Web site interface with a high social presence, such as the perception that there is a medium of communication that represents personal, sociable, and sensitive human contact (Gefen & Straub, 2004).

Social presence has been defined as the extent to which a medium allows users to experience others as being psychologically present (Fulk, Schmitz, & Power, 1987). Social presence theory by Short, Williams, and Christie (1976) describes how the social context affects medium use; they see social presence as a quality inherent in a communication medium. Some researchers characterize the social presence of a medium as its capacity to transmit information about facial expressions, posture, dress, and non-verbal cues (Short et al., 1976). Others focus on the psychological connection, where social presence is concerned with warmth and whether it conveys a feeling of human contact or sociability (Yoo & Alavi, 2001). While others still, focus on its close relationship to information richness theory (Straub, 1994), which concentrates on the interactivity of the media. Related to media information richness theory (Straub, 1994), social presence theory considers the extent to which a medium enables a communicator to experience communication partners as being psychologically present (Short et al., 1976). High social presence is typically found in face-to-face communication. However, medium richness can vary depending on circumstances (Zmud, Lind, & Young, 1990).

A significant difference between online and off-line shopping environments is that the latter involve a wide range of social interactions with humans through multiple sensory channels. Online shopping, on the other hand, primarily tends towards reducing the users’ affective or emotional factors through functional and performance based Web site design. As such, online shopping may be viewed as lacking human sociability, since it is more impersonal and automated than traditional offline commerce. Social responses to computer technology (SRCT) research paradigm proposes that individual interactions with computers and other communication technologies are fundamentally social and natural, and people interact with computers and mediated stimuli using the same social attitudes and behaviors that they apply to other people. The computers are social actors (CASA) paradigm (Nass, , Moon, & Carney, 1999) initiated this field of research and uses theories and experiments derived from psychology, sociology, and communication to develop and validate theories. Granted, online shopping Web sites typically involve no actual interaction with other people, however, Web site interface features have been suggested to help impact the perception of social presence cues, also known as interpersonal cues (discussed in more detail in the next section), that can be embedded in different ways. Higher perceived social presence cues in a Web site may increase online trust through their effect on increased electronic communication since communication is a necessary element of constructive interaction (Gefen & Straub, 2004). Trust may increase when the trusted party shows behavior or other indicators in accordance with one’s expectations; the perception that the vendor is embodying a high degree of social presence cues in the Web site should increase consumer trust to the degree that such indications are expected (Luhmann, 1979).
Social presence cues can convey a sense of personal, sociable, and sensitive human contact, so too should multimedia Web sites. Indeed, recently many online shopping Web sites have used interface features and embedded social presence cues. Advertising research has long relied on imagery of friendly faces to build a positive attitude towards products.

**Social Presence Cues (Interpersonal Cues)**

Social presence cues, also known as interpersonal cues, are the signals that make a person aware of the presence of other people (Short et al., 1976). They are non-verbal and para-verbal communication methods and suggest participant appearance each time a person interacts with another. It may be intentional or unintentional and it is part of the rapid stream of communication that passes between two interacting individuals. Although there are a number of definitions of social presence cues (interpersonal cues), in the broadest sense it is communication that transcends the bare elements of the written or spoken word. The interpretation of this communication has been shown to have a central effect on participants’ perceptions of the event and can give information about individuals’ backgrounds, motivations, and emotional attitudes. They augment spoken messages by helping people express their feelings or thoughts through the use of their bodies, their facial expressions, and their tone of voice and so on (Mehrabian & Epstein, 1972). Psychological studies have concluded that more than 65% of the information exchanged during a face-to-face interaction is expressed through non-verbal means. Fromkin & Rodman (1983) suggest that up to 90% of the communicative process takes place non-verbally. The characteristics of the individuals involved and their response to coding and decoding signals govern the role of social presence cues in any exchange. Three specific variables can be identified as impacting on the nature of non-verbal communication during the exchange; gender, culture, and personal traits. Social presence cues (interpersonal cues) are of relevance in the debate on trust as they can be interpreted as signals for trustworthiness. A trustor can form an impression of perceived trustworthiness of a trustee from the interpersonal cues he/she perceives in a face-to-face situation; it works as inferences (Steinbruck, Schaumburg, Kruger, & Duda, 2002). Interpersonal cues lead to an instant impression formation and thus have an immediate impact on affective trust. The different types of cues differ in their reliability for trust assessments, and in how they are affected by transmission over media (Riegelsberger et al., 2005).

There are many classifications of interpersonal cues (Hinton, 1993); the most common types are:

1. **Paralanguage**: The vocal cues that accompany spoken language
2. **Kinesics**: Body movements
3. **Oculuses**: Eye behavior
4. **Appearance/artifacts
5. **Proxemics**: The non-verbal study of space and distance
6. **Haptics**: The non-verbal communication study of touch
7. **Olfactics**: The non-verbal communication study of smell
8. **Chronomics**: The non-verbal communication study of time
9. **Facial expressions

**Online Trust and Media Cue**

Existing technologies allow for various representations of interpersonal cues that are embodied in different kinds of media cues (e.g., photo, audio, video, embodied agent) to be integrated into one platform. There are many theoretical concepts that classify and explain such media effects. The most common are social presence (Short et al.,
The reduction in the number of interpersonal cues when interacting online is seen as one of the reasons for the lack of trust online (Shneiderman, 2000). This assumption is well supported by many studies (e.g., Hassanein & Head, 2004). These studies found that richer representations result in higher awareness, preference, and interpersonal focus. The most commonly used example of visual interpersonal cues is the smile—this has been identified as powerful in stimulating immediate affective responses, and can form a basis for affective trust.

Thus, Web site interface elements such as photographs, video clips, or synthetic voices can create some level of social presence that may enhance the affective trust compared to Web site interfaces lacking such elements, and can therefore be taken as an indication of the trustworthiness of the trustee.

There is much research on trust in e-commerce vendors, while in contrast there are very few that have specifically tested the effect of interpersonal cues on trust (Al-Diri, Hobbs, & Qahwaji, 2006). Some of the existing studies focused on the effect of synthetic interpersonal cues in avatars (embodied agents); others investigated the effect of displaying facial photos of humans on e-commerce sites. All of these studies tested users’ trust either in the form of quantitative questionnaires or with qualitative interviews, using two mock-ups of an e-commerce Web site, one with and one without a photo or avatar (e.g., Bickmore & Picard, 2005).

Video

In general, video is considered to be a rich media cue and thus is considered to afford high social presence cues as it transmits many visual and audio cues.

In a study by Swerts, Krahmer, Barkhuysen, and Van de Laar (2003), on the effect of social presence cues on the detection of speaker uncertainty, they found the best discriminative ability for video with audio, compared to video only and audio only.

Brosig, Ockenfels, and Weimann (2002) found that video with audio reached levels of cooperation that were close to those reached in face-to-face communication, even though they were reached after a longer time than in face-to-face communication.

Horn, Olson, and Karasik (2002) in a study on lie detection in job interviews found a better performance in discriminating lies from truthful statements in high quality video than in audio only.

The study also compared low resolution video and low frame rate video. They found that low resolution video, which suppresses detailed facial cues gives a performance as high-quality video and good performance in lie detection. Horn et al. (2002) attribute this result to the reduction in truth bias in the absence of recognizable facial cues and the fact that the presence of facial cues may lead to a trusting reaction.

A study by Van Mulken, Andre, and Müller (1999) investigated trust in advice. This study varied the representation of an advisor (video, embodied agent, audio, and text only) and the quality of the advice. Hence, the effect of media representation could be compared to the effect of advice quality. The study found a preference for high quality advice in all representations, but only insignificant indication of a positive effect for video on the behavioral measures.

Photos

Research on the use of personal photos in Web sites are very few and contradictory, with some studies finding such images to be a positive cue (Fogg, 2002; Nielsen, 1996; Steinbruck et al.,
2002), while others found them to be neutral (Riegelsberger, Sasse, & McCarthy, 2002). Urban, Fareena, and Qualls (1999) found that screensized facial photographs of shopping assistants embedded into a shopping Web site interface led to a wide range of reactions as some users liked it, while others considered it unnecessary. In the Fogg et al. (2001) study on the credibility of online news articles, they found that photos of authors increased credibility. Riegelsberger, Sasse, and McCarthy (2003) found that virtual re-embedding had a positive effect on user trust for medium experienced shoppers. Highly experienced Internet users, as well as consumers with a high level of distrust towards online-vendors, benefited little from the provision of social cues in the interface (Riegelsberger & Sasse, 2002). Steinbruck et al. (2002), in an experimental study, investigated whether adding a photo of an employee to the home page of an online-banking site increased user trust in the bank—they found a positive effect on trust. As a result of the foregoing it was hypothesized that:

H-1: Subjects differ significantly on their rating of trust belief and trust intention across vendor’s Web sites.

H-2:

H-2-a. The first rating of a vendor’s Web sites trustworthiness will result for those presenting video clips.

H-2-b. The second rating of a vendor’s Web sites trustworthiness will result for those presenting photos.

H-2-c. The third rating of a vendor’s Web sites trustworthiness will result for those without photos.

Web Site Design and Culture

Understanding how to build trust for diverse consumers in electronic markets is imperative (Grewal, Munger, Iyer, & Levy, 2003). Culture has implications in Internet settings as well and is considered to influence consumer trust (Jarvenpaa & Tractinsky, 1999), Internet diffusion (Ferle, Edwards, & Mizuno, 2002), Internetmarketing (Tian & Emery, 2002), Web site development (Kang & Corbitt, 2001), and Web interface acceptance and preferences. Despite an anticipated large number of consumers from multiple cultures, few studies have examined the Web preferences of users in terms of design characteristics across cultures (Chen & Dhillon, 2003). Cyr and Trevor-Smith (2004) found statistically significant characteristics in Web design for municipal Web sites across cultures. Further, Simon (2001) examined cultural differences related to Web site satisfaction among the residents of Asia, Europe, Latin and South America, and North America based on Hofstede’s model and found different preferences for colors and navigation. Sun’s (2001) exploratory study examining cultural markers focused on language, pictures and graphics, colors, and page layout and found culture to be an important design consideration that increases usability of multilingual Web pages. All of these studies point to one general conclusion—Web sites need to be adapted to the different cultures of the targeted consumers. Singh, Xhao, and Hu, (2003) concluded: “The web is not a culturally neutral medium.” Studies like that of Hillier (2003) emphasized that “building trust on the web requires user interface characteristics appropriate for culturally diverse audiences.” An effective Web site design can engage and attract online consumers’ trustworthiness (Fogg et al., 2001), which is also considered central to trust development (Egger, 2001).
Cultural Photo as a Symbol

Symbols are important elements denoting culture (Marcus & Gould, 2000). One important form of symbolism is multimedia relating to culture, which few researchers have so far examined. Such multimedia elements might include streaming video, sound, and animation. So, the aim was to identify whether the strategy of adding a facial photo representing the users’ culture is beneficial when tested and can bias the trustworthiness of vendor Web sites. Hence the following additional research hypotheses were proposed:

H-3: Across Web sites including human portraits there will be significant statistical differences in trustworthiness between Web sites with local social presence and Web sites with foreign social presence.

H-4: Saudi subjects will trust a Web site with Saudi social presence (photo) more than a Web site with Western social presence (photo).

EXPERIMENTAL METHODOLOGY

This study was designed as a one-factor experiment manipulating three levels of Web site social presence cues (or interpersonal cues). Each of the four specially designed Web sites displayed the same products but each represented different vendors. Only the media cues were manipulated on the sites. Thus, the study attempted to investigate and examine the effects of the interpersonal cues that can be manipulated by facial photographs, video clips, and culture as control variables, which used photos of Saudi and Western people when forming the initial trust toward online vendors. In addition the study set out to measure some auxiliary parameters.

Experiment Participants

The experiment was conducted in an IT training institute in Riyadh, Kingdom of Saudi Arabia (KSA). The number of participants totaled 72 students. Since online consumers are generally younger and more highly educated than conventional customers, student samples are close to the online consumer population (McKnight et al., 2002) and therefore are representative of online shoppers (Gefen & Straub, 2004).

Experiment Material

An initial survey on the most popular online products was carried out and found that laptops came first in popularity. The stimulus experiment material used in this research was carefully selected, having noted that previous studies in online trust used mock-ups of shopping sites to test the effect of interpersonal cues (e.g., Riegelsberger et al., 2002; Steinbruck et al., 2002). This experiment used semi-functional copies of existing vendors’ sites, chosen in consultation with the four most famous reviewer business sites; BizRate.com, ResellerRating.com, Price Grabber.com, and Epinion.com to facilitate the task of rating online shopping sites. These services aggregate feedback from customers of e-commerce vendors based on post service and handling of privacy and security, which represents an aspect of vendor trustworthiness. Western shopping sites were selected as they constitute a realistic scenario with relatively high risk, due to the vendor and the users being in two different countries. The selection was based on the rating of high trustworthiness of the vendors and the number of reviewers of the selected site. Also we made a usability test for the four vendors’ Web sites to check the usability index for each Web site and to make sure that all selected Web sites had almost the same usability criteria. To do that, we asked five participants to use the check list developed by Keevil (1998) to measure the us-
ability index of the four Web sites. The checklist has over 200 questions in five categories, and even though it is not very recent, it is still valid and robust. The results showed that the four Web sites had almost the same usability index.

Semi-functional copies of the Web site including the home page and some subsequent layers depending on the available links in each layer were designed so that participants were able to browse and search general information about the site. Also any certification or reputation seals that were present on some pages were removed.

Photos and the video clip were selected by five professionals in computing and business, who were asked to rate the photos and select the most appropriate based on a realistic image of a customer service representative. The media cues (photo or video clip) were placed in attractive places in the first page of the site showing the selected product (without deleting or hiding anything from the page itself). This page was connected to the entire Web site so the subject could browse and search the site. In addition, each media cue was presented into each vendor Web site based on a predefined Web sites display scenario (see Figure 1).

Data Collection

The research methodology of studies on trust in e-commerce can be categorized into qualitative approaches, using either semi-structured interviews (e.g., Egger, 2001), or qualitative interviews in conjunction with user evaluation.
sessions (e.g., Witkowski, Neville, & Pitt, 2003). The second category entails trust questionnaires in an experiment environment, in this approach the majority of studies used an experiential survey, that is, participants were asked to navigate to a specified or self-selected Internet company and had to perform several predefined tasks and afterwards report on their impressions by filling out a questionnaire (e.g., Jarvenpaa et al., 2000). A second group of studies that used this approach, applied a basic survey approach, that is, subjects were administered a questionnaire or they were pointed to an online-questionnaire form without previously visiting any e-commerce Web site (e.g., Bhattacherjee, 2002). The third category involves social dilemma games, studies in this category measure players’ rate of cooperation and defection to infer media effects on trust and trustworthy behavior (e.g., Brosig et al., 2002).

From a clear review of these approaches, there are distinct advantages and disadvantages. For this study it was decided to use the hypothesis-testing laboratory and experimental quantitative approach. This approach allows a high level of experimental control for independent variables while keeping resource requirements relatively low.

All experimental tasks during this research experiment were performed in a computer laboratory. The research instrument to measure the constructs of interest was developed by adapting existing measures from the literature to the current research context. All items were scored on a five-point, Likert-type scale ranging from (1) strongly disagree to (5) strongly agree (see Appendix 1). As the experiment was conducted in an Arabic speaking country, the questionnaire, originally written in English, was translated into Arabic by a bilingual person whose native language is Arabic. The Arabic questionnaire was then translated back into English by another bilingual person. This English version was then compared with the original version and no items were found to deviate significantly in terms of language. This process was conducted not only because it can prevent any distortions in meaning across cultures, but also because it can enhance the translation quality. The study question items consisted of five sections. The first section included the basic demographic characteristics such as age, gender, education level, and Internet experience. In the second section, respondents were asked to answer questions on their online purchasing experience. In the third section, respondents were tested for their trust propensity or disposition to trust an individual trait, defined as a “general willingness based on extended socialization to depend on others” (McKnight & Chervany, 2001). The questionnaire items were selected based on their high reliability and validity to discriminate the construct and achieve a high Cronbach alpha value; this was adapted from Teo and Liu (2005) and Gefen (2000). In the fourth section, respondents were asked to answer a question adapted from McKnight et al. (2002) regarding their belief towards the Internet as a new medium or as a new environment (i.e., how confident they feel in the Internet system). This item was adapted based on high value of reliability and validity—its Cronbach alpha value was 94%. The fifth section tested respondents on their trust and trust intentions toward the e-commerce vendor. The questionnaire items reflected the most common trust belief dimensions, which are ability, integrity, and benevolence. Also it measured trust intentions—intentions to engage in trust-related behaviors with the Web vendor. All items in this section were adapted from Kammerer (2000), since his questionnaire items effectively discriminate trustworthy from less trustworthy online vendors. Also, Kammerer reports excellent reliability scores with Cronbach alpha values of 0.97%.

Experimental Procedure and Tasks

Previous online trust research has been criticized for relying on measuring trust without inducing any form of risk (Riegelsberger et al., 2003),
which represents an important key related to trust. To overcome these criticisms, this study induced financial risk in a laboratory situation. While it does not fully represent a real-world risk, nevertheless, it allowed combining a laboratory setting with some element of real-world risk by informing participants that the experiment Web site trustworthiness had been assessed and rated by independent business reviewer sites and one of their tasks was to identify the trustworthiness of each shopping site; those whose rating matched the real trustworthiness rate would be entered into a lucky draw with prizes such as a laptop to be offered in a random draw conducted at the end of the study. In this way we induced a slightly higher level of risk thus, increasing the realism of the experiment and encouraging participation.

At the beginning of the experiment a brief introduction and the total estimated time that it would take were given. Then participants started the first part of the experiment by opening the first experiment page, which includes the experiment objectives and instructions, then they started filling out sets of the first four questionnaires sections, which extracted some demographic characteristics, online purchasing experience, disposition to trust, and their belief towards the Internet as a new medium or as a new environment. **In the second part each subject was asked** to look at four Web sites and browse them. This involved looking at the site and then evaluating the e-commerce vendor using the online vendor trust questionnaire. This process was repeated for all four Web sites in the set. However, in order to avoid predictability and the effect of learning by subjects within the scenarios (the possibility that participants exposed to a particular interface or task would do better the next time they were asked to use it), we defined each media cue to a particular vendor Web site since they had almost the same usability criteria. Also, within the experiment software we included a software module with the capability to control the order display scenario of the four Web sites so that different subjects would receive each of the four experimental Web sites in a different order.

When subjects finished, they were asked to move on to the third part which comprised one task. In the first task each subject was asked to search the first Web site seen and imagine that they had enough money to buy a laptop that would serve their needs for the next two years. The subjects were required to indicate the model, price, and certain product specifications of their final laptop choice. Once the participants had found their products and completed their responses on the instruction form, they were asked to fill out the vendor trust questionnaire again.

**DATA ANALYSIS**

All data analysis was conducted using the SPSS Windows software package. A total of 72 subjects participated in this study, all male and between the ages of 18-25 and 26-35 respectively; most (79.2%) held bachelor degrees. As expected, this group was Internet-savvy with over 39% of the respondents spending between 6 and 10 hours online per week. On average, the majority made at least one online purchase per week and 28% of the respondents spent 2000SR (1$=3.75SR) or more per online purchase.

**Testing the Research Hypotheses**

The experiment analysis is based on the well-known statistical test of Friedman. This test is an alternative to the repeated measures analysis of variance (ANOVA), when the assumption of normality or equality of variance is not met. This, like many non-parametric tests, uses the ranks of the data rather than their raw values to calculate the statistic.

This test was chosen because it is suitable for comparing variant vendor Web sites embedded with different media cues when the relative data sample has some characteristics also presented in
this context, including the following ones (Peter & Smeeton, 2001; Sheskin, 2004):

- The experimental data is not parametric, that is, the dependent variables are not normally distributed.
- The number of compared Web sites is two and more. In our context, we typically compared several vendors Web sites embedded with different media cues.
- The same subjects were used to observe different manipulation in vendors’ Web sites.

To test the first hypothesis (H-1), the Friedman test was used. This test was computed for trust belief and trust intention of all Web sites to see if there were any significant statistical differences between the subjects’ answers with regard to the trustworthiness of the four Web sites. Two types of data were used in the analysis, the first was the mean value of each subject answer, and the second was the actual subject answer regarding each questionnaire item. Results showed the subjects differed significantly on their rating of their initial trust and trust intention regarding the four vendors’ Web sites owing to the overall statistical significance (P =.000) at 0.05 levels using the two kinds of data; so, the first hypothesis was supported (see Table 1).

In order to test the second hypothesis (H-2-a, b, c) it was necessary to conduct comparisons contrasting specific conditions with one another. Wilcoxon matched-pairs signed-ranks test was recommended by Sheskin (2004) and Peter and Smeeton (2001) for conducting the three pairwise comparisons (video clip, photo, no photo) with regard to trust belief and trust intention. Table 2 contains the results.

| Table 1. Friedman test for trust belief and trust intention between the four Web sites |
|---------------------------------|----------------|----------------|----------------|
|                                | Trust Belief    | Trust Intention |
|                                | Ability | Integrity | Benevolence | Ability | Integrity | Benevolence |
| N                               | 72      | 72         | 72          | 72      |
| Friedman test Sig at 5%        | Yes     | Yes        | Yes         | Yes     |

| Table 2. Results of Wilcoxon signed rank test |
|---------------------------------|----------------|
| Web site                        | Wilcoxon Signed Rank Test Asymp.Sig |
| With photo                      | Trust Belief   | Rank | Yes   |
| With video clip                 | Trust Belief   | 2    | Yes   |
| With no photo                   | Trust Belief   | 3    | Yes   |
| With video clip                 | Trust Intention| 2    | No    |
| With no photo                   | Trust Intention| 2    | No    |
After reviewing the three analyses, it was found that all pairwise comparisons were significant, with the exception of the video-photo Web site comparison in trust intention. Subjects rated the trust belief and trust intention for the photo Web site the highest, the video clip Web site next highest, and the no photo Web site as the lowest. Thus, the second hypothesis was only partially supported, since the vendor Web site utilizing the video clip came second rather than the expected first position. A possible explanation for this unexpected result is that the video clip was not recorded to professional standards.

The same procedure followed when testing the first, and the second hypothesis was also used to test the H-3 and the H-4 hypotheses, but in this case between two vendor Web sites only, (the Web site with the Saudi photo and the Web site with the Western photo). The Friedman matched samples results are shown in Table 3. The analysis showed that subjects differed significantly on their rating of their trust belief (ability and integrity, although not for benevolence dimension), and also for trust intention in respect of the two vendor Web sites (overall statistical significance of p=.000 at 0.05 levels using the two kinds of data); so the third hypothesis was fully supported (see Table 3).

With respect to the fourth hypothesis (H-4), the Wilcoxon matched-pairs signed-ranks test was conducted for comparison, contrasting specific conditions with one another with regard to trust belief and trust intention. Table 4 contains the results.

From Wilcoxon test analyses, it can be said that pairwise comparison was significant. Subjects rated the trust belief and trust intention for the

---

**Table 3. Friedman test for trust belief and trust intention between the two Web sites**

<table>
<thead>
<tr>
<th></th>
<th>Trust Belief</th>
<th>Trust Intention</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ability</td>
<td>Integrity</td>
</tr>
<tr>
<td>N</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>Friedman test Sig at 5%</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Table 4. Results of Wilcoxon signed rank test**

<table>
<thead>
<tr>
<th>Web site</th>
<th>Trust type</th>
<th>Rank</th>
<th>Wilcoxon Signed Rank Test Asymp.Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Saudi photo</td>
<td>Trust Belief</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Trust Intention</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>With Western photo</td>
<td>Trust Belief</td>
<td>2</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Trust Intention</td>
<td>2</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Web site using a Saudi photo as the highest; the Web site with a Western photo next. So the fourth hypothesis was supported.

**Analysis of the Exploration Depth**

In terms of investigating the factor depth of a users’ exploration of a site (exploration depth) which influences the effect of media-subject on user trust (Research Goal 3), a superficial exploration of the home page and additional pages with no task browsing was compared to a subsequent in-depth exploration of a vendor’s site, beyond the home page with predefined task browsing. Hence, each participant in this scenario saw the first vendor’s site twice: first only the home page and the additional pages that were linked from it, but with a general browsing task (superficial exploration); then the home page and additional pages, but with the specific task of imaginary purchasing—an in-depth exploration. The introduction of the exploration depth was used for several reasons. Most trust models indicated that browsing time is an essential factor affecting trust. Initial trust in the first encounter depends on factors other than behavioral trust that emerges from a long-standing relationship. Applying this consideration to the initial trust that can be measured in a laboratory experiment, it can be proposed that the level of trust will be based upon factors other than initial trust after a detailed exploration of the site. A media cue might have a strong effect based on a quick look, but might lose its significance once more information (e.g., from a site’s security and privacy policies) is obtainable. On the other hand, it can be expected that the first impression (e.g., influenced by a media cue) can color further information processing and thus have a long-term effect (Baron & Byrne, 2004).

**Table 5. Friedman test for trust belief and trust intention between the four Web sites**

<table>
<thead>
<tr>
<th>Ability</th>
<th>Integrity</th>
<th>Benevolence</th>
<th>Trust Belief</th>
<th>Trust Intention</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>Yes</td>
</tr>
<tr>
<td>Friedman test Sig at 5%</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Table 6. Friedman test for trust belief and trust intention between the two Web sites**

<table>
<thead>
<tr>
<th>Ability</th>
<th>Integrity</th>
<th>Benevolence</th>
<th>Trust Belief</th>
<th>Trust Intention</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>Yes</td>
</tr>
<tr>
<td>Friedman test Sig at 5%</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
The Human Face of E-Business

The same analysis techniques were used in investigating the data. The analysis confirmed the same finding, that there was a statistical significance between trust of the four vendors’ Web sites, where the Web sites with a facial photo were ranked first and then those with a video clip. Also the vendor Web site with a Saudi photo was trusted more (statistically significant) than that with a Western photo (see Tables 5, 6, 7, and 8).

Table 7. Results of Wilcoxon signed rank test

<table>
<thead>
<tr>
<th>Web site</th>
<th>Trust type</th>
<th>Rank</th>
<th>Wilcoxon Signed Rank Test</th>
<th>Asymp.Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>With photo</td>
<td>Trust Belief</td>
<td>1</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trust Intention</td>
<td>1</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>With video clip</td>
<td>Trust Belief</td>
<td>2</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trust Intention</td>
<td>2</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>With no photo</td>
<td>Trust Belief</td>
<td>3</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trust Intention</td>
<td>3</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

Table 8. Results of Wilcoxon signed rank test

<table>
<thead>
<tr>
<th>Web site</th>
<th>Trust type</th>
<th>Rank</th>
<th>Wilcoxon Signed Rank Test</th>
<th>Asymp.Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Saudi photo</td>
<td>Trust Belief</td>
<td>1</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trust Intention</td>
<td>1</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>With Western photo</td>
<td>Trust Belief</td>
<td>2</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trust Intention</td>
<td>2</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

Finally, many nonparametric correlation tests were conducted to see whether there were any significant relationships between the trust belief; trust intention and participants’ trust disposition; system assurance; age; education level; or Internet usage. Results showed no statistically significant differences between any of these variables.

DISCUSSION

This experiment investigated initial trust in e-commerce vendors. It explored the effect of adding a media cue (video clip or facial photo) to the home page of an e-commerce vendor and the effect of interpersonal cues that are implicit in media cues, on users’ trust in that vendor. It further studied the effect of the culture represented by a facial photo of a Western or Saudi man on users’ trust in that vendor. The factor depth of a users’ exploration of a site (exploration depth) was found to influence the effect of media cue on user trust. On a methodological level, the study aimed to overcome part of the limitations of conventionally used trust questionnaires conducted under conditions of no
financial risk by introducing trust questionnaires that were elicited under financial risk.

The analyses found that there were significant differences in vendors’ trustworthiness. This supports earlier research by Fogg et al. (2001) and Steinbrueck and Schaumburg (2002), and provides support for the first hypothesis. It was found that media cues can manipulate and bias users’ judgments towards vendors’ trustworthiness based upon inspection of surface cues. Thus, this experiment demonstrated that users can perceive the trustworthiness of a vendor from the surface cues that are present in a media cue displayed as part of the user interface.

Results from the Wilcoxon Signed Ranks test for responses given by subjects on the vendors’ trust belief and trust intention showed favor for the Saudi photo first, with the Western video clip as second and the Western photo as third in the superficial exploration task, while after in-depth exploration participants showed more trusting responses to the media cue vendors’ in terms of trust, and in their declared measures of preference.

An interesting and important result is that the effect of trust was improved once users explored vendors’ sites in greater depth, once they had looked further for a specific task in pages beyond the home page. It can be interpreted that the cues that carried reliable information about a vendor’s trustworthiness were perceived, and media cues could easily influence this decision. Hence, a media cue has a long-term effect on user trust and it biases the perception of other site elements (Baron & Byrne, 2004). Rather, the additional information perceived on additional pages improved the relative impact of the media cue.

In contrast to earlier studies on the effect of photos on e-commerce sites, this experiment investigated the effect of two different cultural photos (Western man and Saudi man) across two different e-commerce sites. This experiment found a bias for the presence of the Saudi photo over the Western photo in terms of trust measures in the superficial exploration task, in the-depth exploration task, and in terms of preference.

CONCLUSION

This experiment addressed the main research goals of the study. First, it investigated a media cue’s (facial photo or video clip) ability to influence user trust in respect of e-commerce vendors, based on surface cues that were implemented in the media cue. This goal is of high relevance, because this strategy is now being used by e-commerce vendors (especially facial photographs, but not yet video) in an attempt to increase trust and attract customers. Secondly, this research tested the effect on user trust of adding a facial photo from two different cultures (Western and Saudi) to an e-commerce vendor’s home page. It thus focused on the symbolic use of interpersonal cues. This goal, despite its importance for the development of trust in e-commerce, has not been addressed in previous research. Thirdly, the experiment was concerned with investigating how time, as an independent variable, represented by the depth of a users’ exploration of a site (exploration depth), influences the effect of a media cue on user trust by comparing a superficial exploration of the home page and additional pages with no task browsing to a subsequent in-depth exploration of a vendor’s site beyond the home page with predefined task browsing. Fourthly, most of the previous studies tested the effects of adding one photo to a mock-up of one e-commerce site. This experiment aimed to overcome this limitation by testing several photos on several semi-functional copies of existing vendors’ sites. In addition, this experiment introduced a method for measuring trust that required participants to make decisions under conditions of financial risk.

This experiment found that media cues in the interface are indeed able to affect a vendor’s trustworthiness based on the surface cues it contains.
A clear picture emerged regarding the effect of photos from different cultures. The experimental results found a positive effect of the media cues in the two stages of exploration—a superficial exploration and an in-depth exploration of a vendors’ Web site. These visceral reactions, however, appear to color the subsequent processing of signals, as their effect improved with the perception of additional signals from an in-depth exploration.

From a methodology point of view, this experiment verified that financial risk can be used in a laboratory experiment to enhance the validity of trust research. Analyzing users’ behavior leads to a relevant outcome to e-commerce vendors since they are concerned with users’ purchase decisions.

With respect to investigating the influence of the auxiliary variables on trust, such as trust disposition, system assurance, age, education level, and Internet usage, the experiment did not find any significant evidence that could support any kind of influence.

Finally, based on the findings of the experiment it is suggested that Web designers and e-commerce vendors should keep in mind the following recommendations when introducing e-commerce applications in Middle Eastern countries in general, and in Saudi Arabia in particular:

There is a significant effect of a media cue in B2C e-commerce Web sites. Users arriving at a site home page will decide, based upon their first impressions, whether to explore the site in more depth or leave it and go instead to a competitor’s site. The positive, attractive impressions of a media cue can thus help e-commerce vendors in the process of converting a visitor to a customer. The findings of this experiment underline the importance of the interface as a communicator of trustworthiness.

In B2C e-commerce applications it is very important to take cultural aspects into consideration when designing an e-commerce Web site. It is expected that when Web sites are appropriate and culturally sensitive, then users will have increased access to content and enhanced user experiences. Selecting photos of customer service representatives that relate to targeted markets represent an easy way to enhance the vendors’ trustworthiness.

REFERENCES


Horn, D. B., Olson, J. S., & Karasik, L. (2002). The effects of spatial and temporal video distor-


Steinbruck, U., Schaumburg, H., Kruger, T., & Duda, S. (2002). *A picture says more than a


Section VII
Critical Issues

This section addresses conceptual and theoretical issues related to the field of human-computer interaction. Within these chapters, the reader is presented with analysis of the most current and relevant conceptual inquires within this growing field of study. Particular chapters address particular algorithms for web personalization, sociotechnical change, and security for ubiquitous computing systems. Overall, contributions within this section ask unique, often theoretical questions related to the study of human-computer interaction and, more often than not, conclude that solutions are both numerous and contradictory.
INTRODUCTION

The individual sense of gendered identity and location are embedded within information technology (IT) usage (Meyrowitz, 1985). Exploring gender in relation to place and IT assists to reveal the impact that cultural knowledge has upon IT usage. This article illustrates the intertwined complex of issues that associate gender and place with IT by examining the currently dominant approaches to research conducted around this topic. The article begins with a presentation of the founding concepts regarding location and the construction of the gendered self, and then discusses investigations of gender, place and IT usage that utilize geographical and bounded constructs and, alternatively the concept of socially occupied space.

BACKGROUND

Situating experience physically—and by implication, locationally—reveals how socio-cultural and socio-technical knowledges play a significant role in the shaping of human action. Maintaining a sensitivity to sociality within IT-oriented research prevents solely technical considerations being interpreted as primary causal factors. Culturally shared knowledge—including media events, physical location, physical sensation and sexual innuendo—is utilized in all IT-enabled exchanges of information. These exchanges serve to strengthen “group” identity and solidarity. Although computer-mediated presence is not a physical environment, the range of possible experiences that can be achieved through engagement with IT-enabled environments are influenced by existing inequalities in the enveloping physical environment. The continual presence of this surrounding physical environment influences knowledge formation to all activities conducted in the IT-enabled environment. IT usage can locate the user—and their actions—simultaneously at many hundreds of machines, creating a physical indeterminacy of presence that challenges the dualisms of physicality, spatiality and the traditional understandings of gender formation (Green,
Owen & Pain, 1994). While gendered examinations of inequality abound (Adam, Howcroft & Richardson, 2002; Wilson, 2004), explorations of the gendered body in relation to technical artifacts are relatively under-represented within IT-based research. By exploring this gap and the intersections of gender, place and technology, it is possible to identify the significance of existing socio-cultural experience and technical knowledge as both barriers and gateways to knowledge acquisition in IT. The conflation of self presence with a “place” makes identity formation a complex phenomenon difficult to isolate or parameterize.

The notion of place and identity are well established in social science disciplines and particularly within geography (Ashkenas, Ulrich, Jick, & Kerr, 1995). “Place” is considered by many (Soja, 1996; Massey, 1994) as primarily a geographic construct providing the individual with a sense of “where” and “when.” However, with few exceptions (see Boland (Jnr), 2001), “modern” organizational forms tend to be analyzed and constructed within traditional and, consequently, physically fixed world views. Predominant discussions (e.g., Gagliardi, 1992) relating to an organization’s physical form reinforce underlying assumptions regarding IT and gender. These positions limit the scope of current analysis in IT research and in relation to gender as self-legitimating arguments that obfuscate critical perspectives.

**IT Artifacts and Human Activity**

The argument presented by this article is indicative of a growing body of research (Gagliardi, 1992) that predominantly combine anthropological and geographical approaches to the examination of human activity and IT artifacts within organizations. Most notably among these approaches is the collection by Gagliardi (1992). These studies critically consider the limits imposed by IT usage on employees’ success in their day-to-day activities. Arguably, the utility and strength of these studies is the parallel considerations of IT artifacts and organizational culture. As Ciborra et al. in Gagliardi (1992) state, “We submit that most recounts overlook the complex interactions between systems as dynamic artifacts and organizational culture.” There is a conscious effort in these works to move away from the causal considerations of computers as tools or control devices towards their role as participatory elements in the organizational environment.

The long history of social constructionist research has tended to look beyond physical form as the primary determinant in the examination of humans and technology. Notions of gender, place, space and even time are seen as social constructs that shape human perspectives of location, boundary and border.

The inclusion of daily lived experience and, specifically, gendered experiences, into the consideration of IT is a significant departure from technologically focused theories associated with implementation and organizational considerations (Ashkenas et al., 1995). Acknowledgement of the significance of cultural roles and everyday life enables an alternative interpretation of the spatiality that is embedded in IT. Geographically focused approaches that consider gender and IT usage, in contrast, utilize physical and observable boundaries (Johnston, 1997), including the spatial limitations imposed by a location. The neglect in explicitly articulating the gendered self of the researcher and the subjects of their research also consequently ignores the socio-cultural aspects of spatial constructions. Examining “place” in relation to the gendered body associates a series of physically bound artifacts, including computers and their intricate ongoing relationship with humans.

**Sense of Place**

Massey (1994, p. 119) describes “place” as a gendered experience that allows individuals to experience “locality” and “inhabitance” along with their own feelings of self-awareness:
So the search after the ‘real’ meanings the ‘self’ in relation to ‘place,’ is the unearthing of heritages and so forth, which is interpreted as being, in part, a response to desired fixity and for security of identity in the middle of all the movement and change. Obtaining a ‘sense of place’, of rootedness, can provide—in this form and on its interpretation—stability and a source of unproblematical identity.

“Unproblematical identity,” which encompasses gender-inclusive identity, is pivotal to the successful inhabitance of IT-enabled places. Inhabitance can be described as the willingness that people have to visit and continue to revisit a space, including IT-enabled environments. This willingness is reflected in the familiarity and closeness that people access and use IT. Anzaldua (1987) offers an example of how familiarity and descriptive closeness can influence relationship and identity formation in relation to “place.” Anzaldua extends her observations of marginalized gendered experiences, providing insight into the intimacy through which she experiences place.

I am a border woman, I grew up between two cultures, the Mexican (with heavy Indian influence) and the Anglo (as a member of a colonized people in our own territory). I have been straddling that Texas-Mexican border, and others all my life. It’s not comfortable territory to live in, this place of contradictions. Hatred, anger and exploitation are the predominant features of this landscape.

However, there have been compensations for this mestiza [a woman of mixed racial ancestry], and certain joys. Living on borders and in margins, keeping intact one’s shifting and multiple identity and integrity, is like trying to swim in a new element, an ‘alien’ element ... [that] has become familiar—never comfortable, not with societies clamour to uphold the old, to rejoin the flock, to go with the herd. No, not comfortable but home. (Anzaldua, 1987, preface)

Massey (1994, p.151) explains that a sense of “place” is established when individuals are able to achieve the combined sense of “locality” and “inhabitance.” The “feeling” of place is achieved through stable identity formation, including and acknowledging one’s gendered identity. In the context of IT-enabled environments, this “feeling” includes identification and familiarity of the IT artifacts being utilized.

Organizations as Bounded Territories and Places

Another way of examining the gendered experience of IT usage within an organization is to employ a spatial perspective. The construction of the organization as a territory, dichotomized between the bounded and unbounded, draws again upon geographic theories as a reference point. Ashkenas et al. (1995) utilize this perspective for their examination of IT. In their example, Ashkenas et al. (1995) emphasize geography in the hierarchical organization that is explored and challenged. The underlying concern in their work is the spatial examination of organizational structure. Spatial understandings of IT simultaneously depict ideas of vastness as well as emptiness. The spatial relativity of objects, such as water and our bodies, are often depicted as extensions of a surface (see Grosz, 1995; Bordo, 1996). The implication implicit in this representation is that extending the surface also incorporates notions of depth and vastness. This understanding of space is also regularly drawn upon to describe and situate human bodies, and therefore, contributes to social constructions of the gendered body (see Grosz, 1995; Bordo, 1996). Density is an aspect of spatial form and is expressed in the popular descriptions of outer space or the depths of oceans. However, spatial notions of depth also incorporate the idea of an inner space. Space conventionally encapsulates a dichotomized position that incorporates the void of an object. Lefebvre’s definition of space is most clearly presented by Molotch (1993).
Space is not simply inherited from nature, or passed on by the dead hand of the past, or autonomously determined by ‘laws’ of spatial geometry as per conventional location theory. Space is produced and reproduced through human intentions, even if unanticipated consequences also develop, and even as space constrains and influences those producing it.

Lefebvre’s understanding of space emphasizes the human requirements for constructing spatial forms. He does this by privileging the subjectivity inherent in the construction of social reality over the object-filled approaches utilized by other spatial analysis and, most particularly, those engaged in studies of geography. These studies are based on Cartesian theory; in other words, the mathematical rules of geometry that extrapolate the socio/temporal relationships of the universe into a manner capable of formulaic manipulation (Woods, 1996). It is Cartesian theory that provides an axis for understanding our “reality” through the duality of time and space. Within this dichotomy, it is time more than space and its impact on work practices, and hence, IT usage that has been, and continues to be, studied (see Boland (Jnr), 2001).

FUTURE TRENDS

Exploring the interaction between humans as gendered beings and objects requires a discussion of “spatiality” and the differing spatial configurations a variety of theorists have presented. “Spatiality” has an association with both time and space. A bodily object is understood in terms of its existence, its connection to “reality” and its mathematical representation (Lefebvre, 1991). The way we spatially position objects in our day-to-day lives is conventionally understood in terms of a human focal point and the relationship between objects and humans. These perspectives are tied closely to Western traditions of Humanism and the absence of the Cartesian object (Jones, 1993). Reality is constructed through existing observable material structures, those “things” that dominate the spaces of daily life. Structures, such as buildings, furniture and other people, can be critically presented as meaning-stabilizing entities (Rosen, Orlikowski, & Schmahmann, 1992). The existence of objects—in “reality”—requires a human physical presence to define its social meaning. When the individual is separated from the social system, it reinforces constructs that assert the presence of an ego “within,” or “inside,” the individual, while society remains “outside” (Elias, 1978). In constrained spatial analysis, such as that of Elias, social systems occupy a boundless space outside the individual and are disjointed from the bounded space defined by the traditionally understood gendered body. A spatial description such as Elias’ is static, placing time and all other social constructs in an arbitrary conceptual opposition to space (Laclau in Massey, 1994). Giddens (1985) as a structuration theorist, and in contrast to Elias’ ideas, interlocks the notions of time and space and, more specifically, time-space compression to interpret the contemporary usage of IT. Utilizing Massey’s theories regarding “place and gender” with spatial theory such as Lefebvre’s allows an analytical shift away from structural and notional time constraints. This shift from the theoretical emphasis found in works such as those of Elias and Giddens is necessary if the conducting of research is to be a determinant of meaning. As Lefebvre (1991) and other spatial theorists argue, it is socially occupied space that can most appropriately be utilized to explore all components of human interaction. In this way, it is arguably one approach that can be drawn upon when interpreting IT usage in a gendered context.
CONCLUSION

Lefebvre (1991, p. 38) defines place as:

... the relationship of local to global; the representation of that relationship; actions and signs; the trivialized spaces of everyday life; and, in opposition to these last, spaces made special by symbolic means as desirable or undesirable, benevolent or malevolent, sanctioned or forbidden to particular groups. We are not concerned here with mental or literary “places,” nor with philosophical topoi [conventions], but with places of a purely political and social kind.

Space in this configuration is the “place” that the individual goes to or chooses not to go to. As an embedded aspect of Western consciousness, space encapsulates all the tensions and oppositions that enable and restrict the individual and the group to interact and conduct their daily lives. Accounting for the gendered place of socially occupied space involves the description of not only the mundane daily visits to the stationary cupboard and canteen but also the quick peek into a fellow worker’s desk drawer to identify those objects associated with gendered identity. Documentary films, ethnographic accounts and the narration of daily activities provide the media for presenting the configurations of place, gender and identity. An understanding of the gritty reality of life and all the limitations and possibilities that unravel in daily life are necessarily presented to the reader once the analysis of a user’s place in relation to IT has been carried out. An example of this theoretical application exists in works, such as Hooper’s (1992) work exploring the production of citizen-bodies in classical Athens, nineteenth-century Paris and Los Angeles through a critical reading of the political emanations from the body of citizen Rodney King. Hooper (1992, p. 50) acts to place disorder at the borderlands of bodies, cities and texts. For example, she writes that:

Body and the body politic, body and social body, body and city, body and citizen-body, are ultimately linked productions ...

The practice of using the individual body as metaphor for the social body, of deploying it as a sign of the health or disease of the social body, develops in the Athenian polis with ideas of the democracy and reason, and continues into the present. Body and city are persistent subjects of a social/civic discourse, of elements and an equally obsessive desire to bring them under control: fear of pollution, contagions, disease, things out of place; desires for controlling and mastering that [become] the spatial practice of enclosing unruly elements within carefully guarded spaces.

The research exploring place, gender and IT echoes Hooper’s (1992) sentiments, as it extends Lefebvre and Massey’s notion of place, gender and identity construction theoretically. “Place” is an avenue from which to explore gender and IT, which, in turn, enables the unraveling of the significance of culture and knowledge acquisition in IT usage. By situating experiences both locationally and, therefore, physically, we are able to show that socio-cultural and socio-technical knowledge play a significant role in shaping the actions of participants. The acknowledgement of “gendered place” in identity formation and IT usage enables sensitivity to sociality not to be subsumed by technical considerations. Considerations of artifacts as they relate to daily experience and identity formation enables an understanding of knowledge acquisition that is IT enabled. The significance of culturally shared knowledge, including gender, physical location and physical sensation, are often neglected when studying users of IT and how they exchange information. Although IT existence is not necessarily tied directly to any physical environment, the range of possible experiences that can be undertaken when engaging with IT are influenced by exist-
ing inequalities in the physical that consequently affect knowledge formation via IT.

REFERENCES


KEY TERMS

**Gendered Body**: Recognition that the physical human body is itself a product of cultural and social forces.

**Gendered Place Experience**: The combined experience of “locality” and “inhabitance” framed within an individual’s own sense of self awareness.
**Knowledge:** Understanding that is actively constructed by the learner; not passively received from the environment.

**Inhabitance:** The willingness of individuals to visit and continue to revisit a space, including IT-enabled environments.

**Place:** Primarily a geographical construct providing the individual with context for the temporal and spatial concepts of “where” and “when.”

**Socio-Cultural Knowledge:** Experiential understanding of the external social world and its forces.

**Spatiality:** The way in which an object is understood in terms of it existence, its connection to “reality” and its mathematical representations.
Chapter 7.2
The Intersection of Gender, Information Technology, and Art

Linda Doyle
University of Dublin, Ireland

Maryann Valiulis
University of Dublin, Ireland

INTRODUCTION

The interdisciplinary field of art and technology is now well established in artistic and academic communities (Wilson, 2001). However, this article will focus on how the combination of technology and art can be used to facilitate the expression of thoughts, the experience of ideas and the explorations of concepts dealing with gender. A research project called the Art of Decision, which focuses on women in decision making, is used as a means of investigating the ways in which creative technologies can illuminate aspects of gender studies.

BACKGROUND

Creative Technologies

In the context of the research presented here, information technology (IT) is defined very broadly as an entire array of mechanical and electronic devices that aid in the storage, retrieval, communication and management of information. It includes all computing technologies and mobile and fixed communication technologies, but it is not restricted to those areas. Smart materials that change attributes on the basis of input stimuli and that can be used to present and display informa-
The Intersection of Gender, Information Technology, and Art

The use of creative technologies with its flexibility, crossing of boundaries, multidisciplinarity and interdisciplinarity lend themselves to feminist inquiry and provide a space to develop feminist...
The Intersection of Gender, Information Technology, and Art

research. At the most basic level, the tools available to us allow material on gender to be presented in a new and exciting way. While this, of course, applies to material of any nature, the use of these techniques in the gender sphere is particularly appropriate.

Gender studies are underpinned by feminist research methodologies. Feminist methodology is interdisciplinary and multidisciplinary, drawing insights from different fields and weaving them together through an understanding of feminist theory. For example, Ramazanoglu’s definition of feminist methodology (Ramazanoglu, 2002) speaks about feminist methodology as being grounded in women’s experience and seeks to analyze connections among ideas, experience and material reality. DeVault, on the other hand, discusses the need for “excavation ... that is to find what had been ignored, censored and suppressed, and to reveal both diversity of actual women's lives and the ideological mechanisms that have made so many of those lives invisible” (DeVault, 1999, p. 30). Reiharz includes in her definition an emphasis on multiplicity of methods and perspectives, of being transdisciplinary, of the effort to create social change, of being inclusive (Reiharz, 1992). Jenkins et al. see “the concept of power as central to feminist research” as well as noting the importance of “how the researcher and the researched have been gendered, sexualized, raced and classed” (Jenkins et al., 2003 p. 2, 4).

Irrespective of the exact definition used, feminist methodologies incorporate the desire to give women an opportunity to tell their stories, express their views and have their voices heard. In essence, it is women-centered. It acknowledges that researcher and researched are “gendered, sexualized, raced and classed,” and both bring these characteristics into the research project (Jenkins, 2002). We consider the interaction of gender with other characteristics of identity as vital to understanding the complexities of research. To summarize very broadly, feminist research methodology: (1) places major emphasis on valuing a variety of viewpoints, (2) is highly concerned with remaining true to the voices of both those who research and are researched, (3) embraces complexity of argument, and (4) incorporates

---

Table 1. Feminist methodology criteria and technology impacts

<table>
<thead>
<tr>
<th>Feminist Methodology Criteria</th>
<th>Technologies of Interest</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capturing of Variety of viewpoints</td>
<td>Communication technologies, wireless networks (mobile communication systems, 2G, 3G, IEEE 802.11, Bluetooth, Zigbee, WiMax etc., ad hoc networks etc)</td>
<td>Communication Applications (web, email, sms, voice, instant messaging, blogs, moblogs etc.)</td>
<td>Communication Interfaces (non traditional input devices, Haptic interfaces etc.)</td>
</tr>
<tr>
<td>Remaining True to the Voices</td>
<td>Devices for capturing, voice, video and text and widely available multimedia applications for the manipulation and or creation of content.</td>
<td>Multimodal means of self-expression reducing the need for verbal or written expression only. Possibility for complete and unedited representation of the voice.</td>
<td>The ability to manipulate and distort and misrepresent data more easily than ever.</td>
</tr>
<tr>
<td>Embracing of Complexity</td>
<td>Applications that allow non-linear presentation of information (hyper text etc.), applications for complex visualization of systems, immersive environments, virtual environments, gaming technologies etc.</td>
<td>The opportunity to deal with complex themes and topics in accessible and interesting ways.</td>
<td>Information overload.</td>
</tr>
</tbody>
</table>
elements of social responsibility and a desire for social change. IT and, in particular in this case, the combination of the technology with artistic practices and methods, can play a major role in the first of these three essential criteria, as shown in Table 1.

As can be seen from that table, the negative potential of the technology is also listed. However, the existence of these very obvious negative factors also has a role to play. Often, traditional science and technology have assumed the air of impartiality and objectivity that gave them the veneer of having produced “authoritative knowledge.” In the current world, this is very much not the case. Exploring gender with creative technologies opens the discussion on how knowledge is produced, authorized and imparted. It helps to problematize questions of objectivity, authority and knowledge production.

The Art of Decision

The ideas introduced in this article are illustrated through an example of a research project undertaken jointly by the Centre for Gender and Women’s Studies and the Department of Electrical Engineering at Trinity College in Dublin, Ireland. This project, known as the Art of Decision, was funded by the Irish government through the Department of Justice, Equality and Law Reform and the European Union. The project aims to bring more women into political decision-making. To show how creative technologies and gender can combine in an effective way, a large interactive multimedia exhibition was specially designed and built. The exhibition took place in a Dublin city center warehouse during May 2005 and was open to the public. The exhibition comprised a series of rooms that immerse visitors in situations that invite them to reconstruct their perceptions of political structures and political involvement as women. Fionnuala Conway (artist and technologist) and Jane Williams (gender studies researcher), both from Trinity College, played a major role in the design of the Art of Decision. Figure 1 shows a schematic of the exhibition. Visitors enter the reception area and wander around the space as they please.

DATAmap

This is a large-scale (47x20 meter) interactive map of Ireland designed to present statistical data on the gender balance on Irish State bodies in more than 70 locations around the country. Users walk on the map and sensors embedded in the map trigger the associated area information on surrounding screens. The DATAmap presents information in a visually interesting and memorable way, depicting the statistics on women and men as pairs of symbols from everyday life – knives and forks, matches and flames, while suggesting a gender for each object. The presentation of the information in this manner highlights the arbitrariness of gender categories and opens discussion on gender categories.

Decisions, Decisions, Decisions

This is a short documentary film where nine people present their perspectives on decision-making. The film is screened in three parts, all three parts running concurrently and in the same location. As the user approaches the three
The Intersection of Gender, Information Technology, and Art

screens, a jumble of voices emerges and it is only by standing in the correct location that sense can be made of the film. In many ways, this project reflects the conflicting nature of decision-making, the “messiness” of the process and the fact that neither men nor women have a monopoly on decision-making procedures.

Finding your Voice

These are simple audio pieces that present unedited stories of two activist women. The stories are presented in two small, intimate spaces as audio installations with lighting design that responds as the stories unfold. The actual voices of the women are heard and their stories are told in their own, unedited words.

Rite of Passage

Images of the visitors to the Rite of Passage are digitally captured and their faces are superimposed on a large digital mural of figures from political life in Ireland, using image signal processing techniques. The power that attaches itself to gender, in this case the masculine gender, is evidence from the dominance of white middle-class men. As women’s faces are superimposed, the balance shifts and we are able to see a more gender-balanced picture.

The Art of Decision Daily Post

This is a giant interactive newspaper projected on a large wall of the room. A headline is automatically pulled at random from an online daily newspaper and displayed as the Art of Decision Daily Post headline and is based on technology by Doyle, Conway and Greene (Doyle et al., 2003). Visitors to the room can text their reactions from their mobile phones to the headline and the reactions appear as text of the newspaper. This is particularly relevant for gender concerns, as it can trigger a spontaneous discussion on current issues important to gender studies.

The VIP Room

The VIP room contains interactive pieces that allow visitors to explore power relations through the manipulation of graphical representations of people’s understanding of power. Drawings from participants are converted to digital format and displayed on the walls of the room. The images can be rotated and explored through control with joysticks. This exhibition illuminates the way in which women see the flow of power and the ways in which the power attaches itself to gendered institutions.

PowerHouse

The PowerHouse is a photographic exhibition presented in a set-designed caricature of a home, its garden and street. Seventy anonymous participants were given disposable cameras and asked to take photos that represent their ideas of power and include comments on the photographs. The visitor is invited to find these photos and comments in the PowerHouse. The research participants’ evaluations of how their attitudes to power changed over the course of the project are also presented over speakers in the PowerHouse. The images can be viewed as an online exhibition at www.imagesofpower.net, where viewers are encouraged to contribute their own comments on the photographs. Using photographs to capture ideas of power provides a different way to capture what power means. Having a camera for a period of time to capture the images creates a persistent alertness to the notion of power. The “voices” of women are heard through the images they chose to take. This type of approach allowed a wide range of women to participate without the need to be able to write skillfully or express themselves eloquently. It also meant that the unedited view of the participants, namely, their photographs, could be presented without interpretation by the researchers.
Rant Room

This is the last exhibit in the journey and is intended as a resting space. Visitors can relax here and text or mail their opinion and comment on the exhibition. The comments board is continuously updated in the space and online during the exhibition so that visitors have the opportunity to send their comment from any remote location via text and e-mail and see it on screen in the space. This exhibition simulates discussion on gender issues. It is an excellent mechanism for allowing voices to be heard in an uncensored manner and for collecting data on views about gender.

FUTURE TRENDS

The ability to stimulate thinking and discussion about gender can increase with the growth in creative technologies. This is particularly important among the younger segment of the population. What was quite striking in the Art of Decision project was the appeal it had to second-level students (students in the 12-18-year age group). They responded positively to the presentation of ideas about gender, gender imbalance and the arbitrariness of gender categories. Thus, joining creative technologies and gender in an interactive manner has the potential to revitalize interest in and discussion about gender among younger cohorts of men and women.

CONCLUSION

Using creative technologies to explore gender enables the researchers to design research projects that cross traditional boundaries and create spaces for women to participate in an active and engaged manner. Using the principles of feminist methodologies focuses the research projects on being women-centered and about doing research in the interests of women. They involve the researched in an active and engaged manner in the evolution of the project. Some allow the participants to change the direction of the research by their input. All stress agency. Moreover, these projects are about raising awareness and consciousness, about presenting information in a new and engaging manner. They tackle issues of power imbalance and work towards bringing new faces into a revitalized feminist debate. When feminist methodology is linked to creative technology, the results are powerful and our understanding of the operation of gender is magnified.

On a theoretical level, the joining of gender, feminist methodologies and creative technologies is most exciting. Creative technologies disrupt the traditional notions of authority and the authorization of knowledge. Because it is technology and technology is allied to science, there is an expectation of authority, validity and objectivity. But creative technology problematizes these expectations and instead focuses on issues of subjectivity, of nonlinear thinking, of multiplicities and imaginings. The questioning of traditional authority characteristic of feminist methodology is given added weight when allied to creative technologies. It reinforces one of the beliefs of feminist methodologies that the authorization of knowledge is not an objective process based on a detached analysis. Rather, it stresses the fact that knowledge and the authorization of knowledge is political and provides the mechanism to reveal that this authorization is raced, classed and sexed.

Moreover, linking creative technology to gender provides opportunities for many, especially those not necessarily comfortable with technology, to not only understand the technological dimension to their project, but to take ownership of it. It calls into question the divide that places technology as a “masculine” tool and allows men and women to participate in this field.

By harnessing the power and potential of creative technologies to feminist methodologies, a richer, stronger, more dynamic understanding of gender may emerge.
REFERENCES


KEY TERMS


Ad Hoc Network: A collection of nodes that form a network on an as-needed basis without the need for any preexisting infrastructure.

Blog: This is the short form of Weblog. A Weblog is a personal journal published on the Web. These journals typically contain informal thoughts of the author in the form of posts or short dated entries in reverse chronological order.

Bluetooth: A technology specification for small form-factor, low-cost, short-range radio links between mobile PCs, mobile phones and other portable devices.

Gender: Gender is primarily defined as the interplay between biology and culture in which definitions of femininities and masculinities are developed and dispersed in accordance with their norms. These definitions are neither unity nor static, so that one can speak of hegemonic masculinities or emphasized femininities.

IEEE 802.11: A wireless local area network standard.

Instant Messaging (IM): A text-based computer conference over the Internet between two or more people who must be online at the same time. When you send an IM, the receiver is instantly notified that he or she has a message.

Moblog: This is similar to a blog except in the case of a mobile Weblog (or moblog)—the content is posted to the Internet from a mobile or portable device, such as a cellular phone.

SMS: Short Message Service, or also known as text. SMS facilitates the sending of text messages on mobile phone systems.
**WiMAX**: A standard for delivering point-to-multipoint broadband wireless access. Specifically, WiMAX is an acronym that stands for Worldwide Interoperability for Microwave Access.

**Zigbee**: This is a proprietary set of high-level communication protocols designed to use small, low-power digital radios based on the IEEE 802.15.4 standard for wireless personal area networking.
Chapter 7.3
Gender-Based Attitudes Toward Technology

Konrad Morgan
University of Bergen, Norway

Madeleine Morgan
University of Bergen, Norway

INTRODUCTION

During the past 30 years of investigation into the ratios of males and females using technology (Harrison, Rainer, & Hochwarter, 1997), there have been consistent reports of males being more positive toward technology and being more likely to adopt the use of new technology on a voluntary basis (Volman, & Van-Eck, 2001). This trend has been reported from early school through adult life, and from diverse geographical sources (Broos, 2005; Heemskerk, Brink, Volman, & Ten-Dam, 2005). Although some scientists have argued that this pattern is changing (Colley & Comber, 2003; Durndell & Thomson, 1997), surveys continue to show an imbalance between the sexes favoring males over females (Colley & Comber; Heemskerk et al., 2005).

The authors consider the consequences of this gender bias to be significant not only in terms of maximizing the whole potential workforce, but also because there is some evidence that males design information- and knowledge-based systems in ways that are different from females, and often these differences favor male users in communication and searching methods. The gender imbalance may become of increasing importance as high-technology industries, such as knowledge engineering and Web commerce, become the normal methods of conducting business throughout the global economy.

BACKGROUND

A scarcity of females in computing can be detected from the earliest levels in many educational systems (Durndell & Haag, 2002). It pervades through all levels of education and into industry (Jackson, Ervin, Gardner, & Schmitt, 2001). This trend could not only pose a threat to the economic growth and stability of the global economy, but it
may also reflect a continuing gender inequality in society.

When we consider the individual differences that occur in humans, then gender, along with age, are often considered among the primary attributes that differentiate people from each other. In comparison to many of the differences such as intelligence, cognitive style, or social grouping, the difference of gender is relatively easy to determine; but like racial origin or social grouping, the topic of gender is often sensitive and highly controversial (Morgan, Brebbia, Sanchez, & Voiskounsky, 2004).

Explanations for these reported gender differences have been varied, but include genetic and hormonal sources (Brosnan, 2004), brain chemistry (Bransford, Brown, & Cocking, 1999), cerebral lateralization (Brosnan), and social roles (Morgan, 2005; Morgan, Gibbs, Macleod, & Morris, 1991; Morgan & Morgan, 2000; Morgan et al., 2004).

FUTURE TRENDS

Reviews of the literature of gender and technology show a consistent trend of male domination in the computing industry and education (Volman & Van-Eck, 2001). Although there appears to be no single reason for this domination, social roles and stereotypes are now thought to be of major importance in shaping education and vocational choices (Durndell & Thomson, 1997; Morgan & Morgan, 2000).

There is a growing body of evidence that suggests that there are strong parental influences on the attitudes and behaviors that we develop in later life. These influences include not only our views on appropriate gender-based behavior (Snyder, Velasquez, Clark, & Means-Christensen, 1997; Tidwell, Witt, 1997), but also our attitudes toward technology and even our self-rated proficiency in using technology (Morgan, 2005; Morgan et al., 1991; Morgan & Morgan, 2000; Morgan et al., 2004). It is the authors’ view that in order to redress the gender imbalance inherent in the technological world of the future, greater emphasis should be placed on parental influence and also that of the educators to encourage females to explore and develop a positive technological attitude. Of equal importance is the establishment of more positive female role models, particularly with regard to the representation of women and technology in the media.

CONCLUSION

As yet, there are no universally accepted explanations for the sex differences found in computing. Broos (2005) conducted a large quantitative analysis of previous studies of the gender divide in ICT attitudes and found, in general, females had more negative attitudes toward computers and the Internet than did men. As we have seen, there is some evidence that social and cultural effects play a large role in gender differences. The mass media also can influence people’s perceptions by the way in which they portray sex roles. Currently, males are usually portrayed as being the predominant users and being in dominant roles in any mixed-sex portrayals (http://www.media-awareness.ca/english/issues/stereotyping/women_and_girls/index.cfm).

One other factor that could explain the lack of females in computing is the harassment of females in the typically male-oriented workplace (Rutter, 1996). This reflects a sexist attitude of viewing females as sex objects rather than human beings or fellow workers. The task of addressing such unfair attitudes, stereotypes, and biased behavior will take considerable time, but it can only be hoped that a day will come when such unfair pressures are removed.
REFERENCES


Morgan, K. (2005). Thirty years of gender and technology research: What have we learnt and how can it be applied? IADIS Virtual Multi Conference on Computer Science and Information Systems (MCCSIS 2005), (pp. 1-6).


**KEY TERMS**

**Cerebral Lateralization:** The theory that the attitude and behavior differences reported as existing between males and females are due to differences in brain specialization. Specifically, left-hemispheric activity is associated with male behaviors, and right-hemispheric activity is associated with female-type behaviors.

**Cognitive Style:** The high-level heuristics or principles that organize mental behavior. Some researchers propose that the natural cognitive style of males and females are not only different, but are responsible for many reported gender differences.

**Gender Roles:** The sets of behaviors and attitudes that a society or culture normally associates with males or females.

**Sexual Harassment:** Unwelcome and unlawful discriminatory behavior or attitudes that are related to sexual or gender roles.

**Technology:** The term technology is used within this article to cover the cluster of technologies that includes all aspects of ICT, networked computers, hypertext, hypermedia, the World Wide Web, and other adjuncts.

**Technology Attitudes:** The views held by an individual with respect to technology that influences not only his or her rate of voluntary use of technology, but also predicts stress levels when faced with situations that involve the use of unfamiliar technology.

*This work was previously published in Encyclopedia of Gender and Information Technology*, edited by Eileen M. Trauth, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
INTRODUCTION

Distance education is defined by six pedagogical elements: (1) physical separation of teacher and learner; (2) regulatory function or influence of an educational organization; (3) media to connect teacher and learner; (4) two-way communication exchange between teacher and learner; (5) individualized pedagogy instead of group focus; and (6) “industrialized” facilitators, entailing less individuated instructors (Keegan, 1980). Distance education technologies include video (videotape, satellite delivery, microwave delivery, broadcast video, and desktop video), computers (e-mail, Web-based courses, video conferences, DVD, and CD-ROM), collaborative activity software (chat, discussion rooms, and white boards), voice/audio technology (telephone, voice mail, audio conferences, audiotapes, and radio), supplemental print material (books, study guides, workbooks, and FAX), mobile technology (laptop computers, PDAs, tablet PCs, and cell phones), and blended-learning combining one or more of these delivery methods together, including face-to-face instruction.

Distance education technologies as tools are situated in the larger context of technological and scientific knowledge, economic institutions, including the property and market institutions of capitalism, and social institutions, such as education, which historically has been unequal and exclusionary due to class structure and the system of gender and racial power relationships (Carroll & Noble, 2001). People barred or deterred from regular access to education in various ways have always been users of distance education technologies, starting with its inception as correspondence course education in the 19th century and continuing today in high tech distance education classes with women comprising the majority of enrolled students (Hansen, 2001; Ossian, Christensen, & Rigby, 1968).
The promise of distance education technologies in the 21st century is for empowerment of students through democratization of knowledge, personalized pedagogy, and convenient access. Despite the promise and the current high enrollments in distance education courses, attrition rate is high in North America and Europe (Carr & Ledwith, 2000; Serwatka, 2005), and this is a concern to educators and social policy makers, who search for reasons to account for the discrepancy between promise and practice. While recognizing men students have high attrition in distance education courses, too, the fact is women comprise the majority of distance technology users. If educators and policy makers hope to use distance education technology to reach female students (and garner the interest of more male students as well), then issues of gender in distance education technology need to be addressed. A female gendered perspective on distance education technology reveals a number of variables that explain women’s disengagement and dissatisfaction with online educational delivery systems. Educators, secondary education institutions, and instructional software designers are some of the groups working to create and implement inclusive, constructivist, and rich multi-media instructional design (McLoughlin, 2001) that will accommodate a wide range of learner needs.

BACKGROUND

Educational technology associated with distance education reflects the evolution of human institutions and inventions, starting from mail delivery and evolving to CD-ROMs and PDAs. From its beginning in the 19th century, distance education stemmed from a scarcity of educational resources in an era with more rigid class and gender definitions than in the 21st century (Carroll & Noble, 2001; Gandle & Katz, 2004). Distance education, first known as correspondence education because course materials and lessons were conducted using the postal service, was directed toward nontraditional, under-privileged and disadvantaged groups who did not have economic, geographical, or social access to education. Correspondence course users included people who were poor, rural, intellectually gifted, and female. In the U.S., distance education was modeled after European correspondence education, such as the program started by Isaac Pitman in Great Britain offering shorthand instruction. In 1892, the University of Chicago established a school of correspondence study that raised the reputation of distance education as an institution in the U.S. Most distance education schools in the United States in this time period were not the caliber of the University of Chicago, however. Often advertising their educational programs on matchbook covers, hence the derogatory appellation matchbook colleges, these schools were considered inferior to traditional education and often were second-rate and even charlatan. With the 20th century advent of new technologies, such as radio and television, however, correspondence education in general took on other forms and its respectability began to improve. As technology continued to evolve, using audio cassettes, video cassettes, cable television, and tapes, so, too, did correspondence education. The modern stage in correspondence education began in the 1970s with the creation of Open University in Great Britain, where print instruction was supplemented with audio-visual technology. Within this time frame, teleconferencing and computers began to be used as instructional mediums. In 1982, the International Council for Correspondence Education evolved into the International Council for Distance Education, and the term correspondence course dissolved.

Today distance education is a legitimate and mainstream pedagogical practice in K-16 education and employee training. Many educators, business leaders, and policy makers are firm in the view that, there is no significant difference in asynchronous learning networks (ALNs) from
University (brick and mortar) settings in terms of learning outcomes (Allen & Seaman, 2003). Currently, in the early 21st century, distance education, along with its various technologies, is an integral part of education, specifically, higher education services, which are projected to reach 160 million students by 2025 (Hansen, 2001).

WOMEN AND DISTANCE EDUCATION TECHNOLOGY

Women have taken advantage of distance education from the 19th century, when, in an era of educational rationing, depending on class, sex, and race, people were denied access to higher education and professional education. (Carroll & Noble, 2001). Now in the 21st century, women are the dominant user group of distance education, although they no longer depend on the postal system for delivery of course material. Instead, the computer is the major distance education technology with instructors utilizing various ALNs or course management systems (CMS), such as WebCT, Desire2Learn, and Blackboard, which allow students to access classes anywhere and any time they have Internet connection. Some instructors create their own CMS by placing material on their own individually designed Web sites, often with a la carte features added, such as discussion boards and chat rooms. The computer skills students need to engage with any of these courses are minimal. Orientation to the online environment is a major requirement, but once this is accomplished, a good deal of activity on the computer entails keying in text and handling the mouse. Although technical difficulties may intrude from time to time, most CMS are user friendly and designed to be problem free.

Despite women’s long history as distance education students and user-friendly CMS, female interaction with distance education technology, specifically computers, is problematic. Generally, women lack confidence with computers and are anxious about encountering technical difficulties, although, ironically, they are more successful in completing online course than men, who tend to have greater confidence and proficiency with computers (Parker, 2003; Serwatka, 2005; Taplin & Jegede, 2001). Attrition in online education is high for women, and many express dissatisfaction with their online learning experience. Men students also experience high levels of attrition and dissatisfaction when it comes to distance education; however, given that the majority of distance education students are women, educators, cognitive scientists, and educational software designers are especially interested in determining why women are not more enthusiastic users of distance education technology. Research suggests a number of gender variables influencing women’s attachment and detachment to distance education technology, all of which contribute to an inequitable and unsatisfactory learning experience: computer socialization, which includes diversity issues; tacit discrimination; androcentric dominance online; lack of support structures for successful course completion; and poorly designed online courses.

COMPUTER SOCIALIZATION

In Western middle class educational settings and home life, gender enculturation in terms of computers is evident at an early age. At home, boys are generally given more access to computers than girls, who instead will often watch males in their households use computers. Toys and games also serve to reinforce this gender divide in computer use, especially since computer games typically have been created and manufactured for boys. Although the educational value of computer games is contested, research indicates gaming does accustom boys to online environments, allowing them to become accustomed to navigating through games and online sites and using discussion boards and forums to locate cheat sheets and
build competencies in gaming, programming, Web site development, and technical troubleshooting. Consequently, before adolescence, boys can have keyboard dexterity and technical proficiency that will be helpful in future coursework and computerized instruction.

The socialized gender divide in computer acclimation becomes more evident in middle school and high school, where girls, even though enrolled in more advanced placement (AP) classes than boys, are not active in computer science classes or clubs. In post-secondary education and college, this pattern persists. Computer science (CS) in college and career choice is a largely male terrain with women earning a low percentage of the bachelor’s degrees awarded in CS in the U.S. Women who do take computer courses in college terminate training much earlier than males; women with CS degrees earn fewer advanced degrees. Socialization, social expectations, and cultural metaphors all guide these choices. (Palmieri, 2004; Rajagopal & Nis, 2003). Girls who are not comfortable with computers grow up to be women who are apprehensive about computer use and this attitude affects their experience of distance education courses, even when they enter or exit courses with the required competency in computer use (Fisher & Margolis, 2002).

**DIVERSITY**

Although women in a gendered world have some historical and social experiences in common, there is no uniform female experience, and any examination of women’s interaction with distance education technology must acknowledge the diversity of women’s lives (Carr & Ledwith, 2000). Socialization with regard to computers and computerized instruction, then, is qualified by differences in each woman’s personality, family socialization, class, and ethnicity, (Hayes & Flannery, 2000). Women who have multiple cultural identities, such as Spanish American and Asian Americans, often experience conflicting demands and expectations from two cultures, adding another variable to gender and distance education technology. Some women, for instance, depending on their ethnic backgrounds, find it difficult to provide critical commentary on written texts or written authorities, which makes distance education assignments requiring critical analysis difficult for them to fulfill. When it comes to working poor and underclass women in the U.S. and first world countries, these women, who often have not completed secondary education, have difficulty utilizing distance education technology because of economics and because electronic culture is unfamiliar in their social and educational milieus.

**TACIT DISCRIMINATION**

A more crucial factor involved with women’s attitude and relationship with distance education technology is tacit discrimination that deters women from computerized instruction and computer use in various ways (Palmieri, 2004). The basic tenet of tacit discrimination, derived from feminist theory, is that men and women lead very different lives in patriarchal societies; gender as a system of social organization and social relationships has effects on learning and knowledge acquisition (Hayes & Flannery, 2000) and makes self-realization and functioning in society more difficult for women. Unlike blatant sexual discrimination, which is now mitigated by statute law, tacit discrimination is unconscious and subliminal in terms of its gender discrimination, making it difficult to recognize and surmount.

One example of tacit discrimination in relation to the gendering of distance education technology is educational institutions not providing safe access to campus computer labs at night, which dissuades women from computerized instruction. Another example of tacit discrimination toward women using distance education technology is the
Gender in Distance Education Technology

failure of institutions and instructors when planning educational programs or designing courses to take into account that school work for many women entails a “third shift” in their lives, along with the first shift of childcare and family responsibilities and the second shift of paid employment (Kramarae, 2000). Although some men students also work a “third-shift” in terms of child care and family responsibilities, statistically women in society still perform a majority of third-shift work. The burden of this third-shift creates not only work overload and increased stress but also anxiety about failure. As a result, many women withdraw from distance education courses or do not successfully complete them.

ANDROCENTRIC DOMINANCE ONLINE

Another factor contributing to women’s dissatisfaction with distance education technology and online instruction is the dominance of men in CS and digital culture. Predominantly, computers and the Internet are male terrains, although women use computers and the Internet for e-mail transmissions, social interaction on discussion boards and forums, and e-commerce (Fisher & Margolis, 2002). Computer hardware, software, tech-speak, and online communication comprise an androcentric or male centered world that does not overtly bar women, but it does not always encourage their participation or engagement. For example, in Internet communication, whether forums, discussion boards, or chat rooms, male dominated speech prevails: aggressive, blunt language; harassment; flaming; and combative, sexualized jargon.

This male patois in electronic culture carries over to distance education technology in chat rooms, discussion boards, and class e-mail transmissions where men tend to dominate communication (Jackson, Ervin, Gardner, & Schmitt, 2001). When women encounter male classmates online with these behaviors, similar to traditional classroom dynamics, the learning environment is perceived as non-nurturing and hostile, causing self-doubt and silence. Research shows that male verbal dominance in online class environments takes many forms, all of which make women feel unwelcome: indicative, imperative, and agonistic verb moods; intentional and unintentional patronizing language; dismissive and hostile comments; and vulgar speech. In this verbal environment, women often feel alienated, consider themselves as outsiders, and are disinclined to further their education using distance education technology (Jackson, et al., 2001; Taplin & Jegede, 2001; Vuorela & Nummenmaa, 2004).

LACK OF SUPPORT STRUCTURES

Still another reason for women’s disengagement from distance education technology is lack of support structures for successful pursuit of educational goals. Lack of support falls in to five categories: institutional, occupational, peer, instructional, and familial, all of which men in distance education courses experience as well. Because distance learning differs from conventional learning primarily in terms of student isolation and self-discipline, it is important for students to have support structures that enable them to successfully engage with the course material (Serwatka, 2005; Taplin & Jegede, 2001). Problems develop for students when these support structures, for various reasons, are not in place.

Lack of institutional support results when campuses operate as if they only have traditional students to service, such as when offices and college services are not open in evening hours. Occupational support systems fail when employers, who are generous with oral encouragement and financial support, do not follow through in terms of flexible work hours that would allow completion of class work. Peer support structures founder when students feel isolated from other students in
the class, both in terms of physical presence and communication lines. Although women typically do well with the self-regulation and independent study required in distance learning, the social isolation from other students is difficult for them (Moore, 1993; Parker, 2003; Serwatka, 2005). Lack of instructional support takes the form of unreceptive, unresponsive instructors who do not respond to e-mail in a timely manner. Close contact with online instructors is another crucial factor in whether women gauge their online experience as meaningful and worthwhile (Serwatka, 2005; Taplin & Jegede, 2001). Finally, lack of familial support also hinders women distance education students. Although women generally like the convenience of working from home and not having to commute, home life does not always nurture educational effort, and psychological or emotional support from family members is absent (Kramarae, 2000).

PEDAGOGICAL SHORTCOMINGS

Questions about women’s attachment to and detachment from distance education technology leads to questions about the quality of the pedagogical theory and design informing the technology as well as the quality of teaching by online instructors. How do designs and practices of distance education technology affect women learners’ experiences? Distance education practitioners suggest five criteria for quality distance education instruction and delivery: facility with delivery and access; learner control over the medium, including adequate skills for human-computer interface; a high degree of interaction, not only between teacher and learner but learner and learners; a rich multimedia learning environments, utilizing audio-visual material; and a strong social presence in the course (Short & Christie, 1976).

Examination of these criteria reveals that, the major distance education technology in use, the computer, is not the problem, even though women experience computer anxiety. Distance education course design, content, and teaching practice, however, do fail to live up to the suggested criteria. Traditional pedagogical frameworks, similar to on-site lecture courses, still predominate in online instruction. Too often, course material is nothing more than classroom text and notes transplanted online (Allen & Seaman, 2003; Lee & Owens, 2000). Additionally, many courses are designed using a middle-class male episteme or knowledge framework, which alienates many students (Moore, 1993; Serwatka, 2005; Tisdell, 2000). Online teaching is also lacking. Many online courses lack strong social presence and have low levels of instructor-student and student-student interaction (Hayes & Flannery, 2000; Moore, 1993). Consequently, men and women students both are disenchanted with their online educational experiences.

FUTURE TRENDS

Educators, instructional designers, and educational institutions are aware of the factors impacting women’s experience with distance education technology and online instruction, and, accordingly, they are making efforts to correct weaknesses and deficiencies. Elementary and middle schools are requiring girls to be computer literate as well as literate in digital culture, and educational and recreational software is being designed for girl students and girl gamers. All these efforts will nurture female computer confidence. Educational institutions are offering affordable day care facilities to accommodate women students with children; providing safe access to computer labs in evening hours, including security guard or buddy-system escort service to cars; and adding evening office hours for access to administrative offices and student counseling.

Educational and feminist theorists, aware of the importance in designing online courses to meet
the needs of the target audiences and the needs of those who are at a disadvantage, whether from gender, ethnicity, or socio-economic background, are designing online classes that promote endogenous learning and knowledge construction for all students (Carr & Ledwith, 2000; Hayes & Flannery, 2000; McLoughlin, 2001; Palmieri, 2004; Tisdell, 2000). Emphasis is placed on universal and inclusive pedagogy in instructional design because it requires instructors and designers to place learners’ needs first in the educational process and design courses for diverse student bodies. Implementing constructivist pedagogy in online instruction, which emphasizes the active participation of the learner, is also viewed as key in pedagogical reform (Weigel, 2001).

Online course designers and educators are also striving to take advantage and make use of rich multimedia software and utilities in online material (Billinghurst, Kato, & Poupyrev, 2001; Knemeyer, 2005; Lee & Owens, 2000), which will stimulate all the senses and intellectual faculties. These designers and educators are aware that more innovation is needed in online instruction, such as an electronic teaching theater that takes advantage of hypertext and multimedia and allows “depth education” (Weigel, 2001) that will engage and challenge students. Learning objects (Los), such as those available through Project Merlot—http://www.merlot.org/Home.po—can also be used to enhance online instruction.

Finally, educational institutions and instructors themselves are acknowledging that, distance education instructors, who are the key to making distance education technology and online instruction a positive experience for students, need training and institutional support on how to design stimulating and challenging online content (Allen & Seaman, 2003; Moore, 1993). Accordingly, faculty development training in online instruction and effective use of distance education technology is now the new standard in post-secondary educational institutions.

**CONCLUSION**

Distance education technology has evolved a long way from the 19th century mail delivery correspondence system to computerized instruction. Statistics indicate women, the dominant user group, successfully complete distance education classes more than men; however, women, as do their male student cohorts, experience high attrition in distance education courses and express dissatisfaction with their educational experiences. There is no single factor responsible for female attrition, nonparticipation and discontentment; the reasons are multidimensional and complex: computer confidence socialization; diversity issues of race, ethnicity, and socioeconomic status, which qualify women’s relationships to digital culture; tacit gender discrimination, such as “third-shift demands of childcare and family responsibilities on women; androcentric dominance online; support network absence; and pedagogical shortcomings in instructional content and instructors themselves. Improving women’s experience with distance education technology begins with awareness of these factors and the acknowledgment that a woman-centered, learner-centered perspective by educational institutions, educators, and instructional designers is needed in order to make effective corrections. Approaching distance education technology from an inclusive perspective, along with applying constructivist pedagogy that is highly interactive and rich in multimedia will humanize and enrich online education for all learners—for men as well as women.

**REFERENCES**

Gender in Distance Education Technology


**KEY TERMS**

**Asynchronous Learning Network (ALN):** An online method of instructional delivery that enables students to have the flexibility of learning at anytime from any location over the Internet.

**Constructivist Pedagogy:** A learning theory that contends learning should be an active process, whereby learners themselves construct new ideas or concepts based upon their current/past knowledge. Constructivist instructors encourage students to discover principles by themselves, and the instructor and student engage in highly interactive, Socratic dialogue.

**Inclusive Pedagogy:** See **Universal Design for Instruction (UDI).**

**K-16:** Refers to kindergarten through four years of college or post-secondary education.

**Learning Objects (Los):** Instructional modules that enable and facilitate the use of educational content online. Due to internationally accepted specifications and standards, Los are interoperable and reusable in various learning environments. (http://itdl.org/Journal/Sep_04/article02.htm)

**Personal Digital Assistant (PDA):** A handheld device that combines computing, telephone/fax, Internet, and networking features. Some PDAs can also react to voice input by using voice recognition technologies.

**Social Presence:** One of the most important factors for student engagement, collaboration, and retention in online learning environments. Social presence is the ability of learners to project themselves as real people and relate to the instructor and class peers socially and emotionally online. Formation of social presence for each student is not automatic; it must be nurtured and supported by the online environment and the instructor.

**Tablet PC:** A compact, portable device that incorporates many features of a laptop computer while also allowing learners to write lecture notes and save them electronically.

**Tech-Speak:** The formal technical vocabulary of programming, computer science, electronics, and other fields connected to hacking. (http://catb.org/~esr/jargon/html/distinctions.html)

**Universal Design for Instruction (UDI)/Inclusive Pedagogy:** An approach to teaching that consists of the proactive design and use of inclusive instructional strategies that benefit a broad range of learners including students with disabilities” (UDI Fact Sheet, 2002). (http://www.facultyware.uconn.edu/home.cfm)

*This work was previously published in Encyclopedia of Gender and Information Technology, edited by E. Trauth, pp. 622-629, copyright 2006 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 7.5

Gender Differences in Information Technology Acceptance

Will Wai-kit Ma
Hong Kong University of Science and Technology, China and Hong Kong Shue Yan College, China

Allan Hio-Kau Yuen
The University of Hong Kong, China

INTRODUCTION

Gender differences in computer use has been always a topic of research interest. The understanding of the patterns among gender, including beliefs, intention and use behavior of IT/IS would provide us a better picture to the process of design and implementation, which gives support to IT/IS success. However, published works explaining why and how beliefs and attitudes varied between different genders are still scarce, yet the topic was of widespread relevance.

We direct our empirical work on user’s beliefs, intention and usage behavior. Gender differences in beliefs would likely make a corresponding impact on the intention to use or not to use computer in the future, and hence, the actual usage pattern. Assumed to be behavioral manifestations of users’ gender differences, we infer these gender differences in the beliefs of computer use from their self-reported intention and usage behavior.

Therefore, we aim to explore the factors affecting the intention and usage behavior; and their corresponding strengths in affecting the intention and usage behavior; in order to suggest effective implementation strategies accordingly.

The research questions of this empirical study are:

1. What are the emergent constructs that drive the intention and usage behavior of computer use?
2. Do users’ beliefs regarding IT/IS usage differ among genders?
3. To what extent do these effects differ?

To address these research questions, we applied technology acceptance model (TAM) to a group of pre-service teachers, and measured their beliefs in using computer to explain the gender differences in their beliefs, intention, and usage of computer. The rest of the article proceeds as
follows. The next section starts with a review on gender and technology. The third section explains the model framework TAM. The fourth section describes the instrument construction and validation. The fifth section reports the model testing results. The final section discusses the thrusts of the study and future trends.

BACKGROUND

There have been findings showing that gender differences in computer acceptance are prevailing. Young (2000) found significant gender differences in computer attitudes of 462 middle and high school students. The male domain scale showed that boys were more likely to have claimed computers as a male area. Thus, higher levels of confidence and, for males, the absence of negative teacher attitudes were associated with greater computer skills. Using TAM as the theoretical framework, Venkatesh & Morris (2000) found that, compared to women, men placed a greater emphasis on perceived usefulness in determining behavioral intention. On the other hand, women weighted perceived ease of use more strongly in determining behavioral intention than men did at earlier time frame. A few more empirical studies showed that gender differences in information technology do exist: Yuen and Ma (2002) found significant gender differences in beliefs while applying the technology acceptance model to a group of pre-service teachers; Houtz and Gupta (2001) found that males generally are more interested in information technology; Gattiker and Nelligan (1988) suggested that there is an association between gender and attitudes of information technology. On the other hand, interestingly, in their study of Australian Women in IT, Hellens and Nielsen (2001) indicated gender and IT were socially constructed as they suggested that cultural differences might be more important than gender alone, “Women of Asian background significantly outnumber all other ethnic female students in Australian IT degree studies” (p. 48). However, whether this applied to teachers is still in doubt and further empirical investigation was in need. Thus, the aim of this article was to explore gender differences in teacher computer acceptance in contrast to the studies in other workplaces. To prepare this article for the “Encyclopedia of Gender and Information Technology”, part of the findings were extracted from a previous article of the authors (Yuen & Ma, 2002).

METHOD

Subjects

The study targeted pre-service teachers who were mostly fresh degree holders, joining the one-year full-time teacher education program (Postgraduate Certificate in Education) at a local university in Hong Kong. According to past experience, majority of these graduates would become teachers and work locally. It was believed that a study to these subjects would provide a good understanding of the pre-service teachers, but also shed light to understand the future computer use of in-service teachers. A summary of the 186 respondents who had successfully completed the survey instrument was listed as seen in Table 1.

Technology Acceptance Model (TAM) and Its Measurement Items

In prior studies, there have been extensive investigations on developing computer attitude scale. Attitude was viewed in a hierarchical manner, including firstly the affective responses to attitude, then the cognitive responses to attitude, and the highest level of conative responses to attitude (Ajzen, 1988). Applying this attitudinal process to computer use, it might explain as: (1) firstly an user heard about computers and tried to evaluate them; (2) then, the user got chance to have hands-on experience with computers and
formed perceptions about computers; and (3) finally, the user reflected his or her attitude on computers through behavioral intention and actual usage behavior. How to measure perceptions would become an important process to predict and explain computer use.

TAM was one of the widely validated and applicable model frameworks to measure perceptions on technology use. It was firstly suggested by Davis, Bagozzi, and Warshaw (1989). TAM suggested that perceived usefulness and perceived ease of use as two fundamental determinants to intention and technology usage. Other empirical tests of the TAM (e.g., Adams, Nelson, and Todd, 1992; Hu, Chau, Liu Sheng, & Tam, 1999) had for the most part, been supportive of the model for the last 20 years. Legris, Ingham, & Collerette (2003) conducted a critical review of the technology acceptance model and confirmed the wide applicability of the model towards a wide range of technologies, organizational contexts, and subject domains.

The use of TAM to investigate student-teachers’ computer acceptance was advantageous because of its well-researched and validated measurement instrument. Specifically, the questionnaire was designed to include five items of perceived usefulness (PU1 to PU5), five items of perceived ease of use (PEOU1 to PEOU5), two items of intention to use (ITU1 to ITU2). All items are measured in a 7-point Likert scale, with 1 indicating strongly disagree and 7 indicating strongly agree. The major measurement items were listed in the appendix. Subjects were also asked to report their self-reported usage. Duration of usage was measured in hours per week and was coded into seven categories, namely, “Less than or equal to 4 hours”; “4 to 6 hours”; “more than 6 to 9 hours”; “more than 9 to 12.5 hours”; “more than 12.5 to 16 hours”; “more than 16 to 20 hours”; and “over 20 hours”. The degree of current usage of computer was measured in a 7-point Likert scale. At the same time, subjects were asked to state demographic data in the first

Table 1. A summary of respondents details

<table>
<thead>
<tr>
<th>Particulars</th>
<th>Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male (24.9%)&lt;br&gt;Female (75.1%)</td>
</tr>
<tr>
<td>Age</td>
<td>Less than 22 (9.8%)&lt;br&gt;22-24 (68.5%)&lt;br&gt;25-27 (12.5%)&lt;br&gt;28-30 (2.7%)&lt;br&gt;Over 30 (6.5%)</td>
</tr>
<tr>
<td>Full-time teaching experience</td>
<td>No teaching experience (87.5%)&lt;br&gt;Less than 1 year (7.6%)&lt;br&gt;1-2 year (3.8%)&lt;br&gt;3-5 year (1.1%)</td>
</tr>
<tr>
<td>Major teaching areas</td>
<td>Art subjects (42.4%)&lt;br&gt;Science subjects (35.5%)&lt;br&gt;Social science subjects (22.1%)</td>
</tr>
<tr>
<td>Access to computers at home</td>
<td>Yes (98.4%)&lt;br&gt;No (1.6%)</td>
</tr>
<tr>
<td>Formal computer training</td>
<td>Not at all (46.3%)&lt;br&gt;1-8 hours (17.9%)&lt;br&gt;9-16 hours (13.6%)&lt;br&gt;17-24 hours (6.0%)&lt;br&gt;25-32 hours (6.5%)&lt;br&gt;33 or above (21.7%)</td>
</tr>
</tbody>
</table>
part of the questionnaire, including gender, age range, major teaching areas, access to computer at home and if there was any previous formal computer training experience.

**Procedure**

Data were collected using a user-reported self-assessment approach. It deemed to be appropriate because of considerable literature support for its use in intention-based studies and being the common method used in TAM research (e.g., Collopy, 1996; Davis, 1989). At the beginning of the semester in October, a total of 282 questionnaires were distributed through the various group representatives. Subjects were asked to return the completed questionnaires to their group representatives within a week’s time, just before they left for their school experience. Group representatives collected the questionnaires and sealed in an envelope and returned to the researcher for collection. 186 questionnaires were collected with return rate 66 percent.

**FINDINGS**

**Summary of the Observed Variables and Scale Validation**

The descriptive statistics of the measurement items were shown in Table 2. It showed that all the items showed generally positive perceptions towards computer use, all mean scores over four. The mean scores ranged from 4.60 to 5.79 while the standard deviations ranged from 1.15 to 1.51. All constructs satisfied the criteria of reliability (alpha > 0.80).

Discriminant validity was demonstrated if an item correlated more highly with items within the same factor than with items in a different factor (Campbell & Fiske, 1959). The inter-item Pearson correlation coefficients showed the discriminant validity where the coefficients of inter-item within each measurement construct were much higher than correlations across constructs. The factor components were then analyzed by a principal component factor analysis with varimax rotation method. The components generated confirmed the corresponding constructs as predicted by the TAM model. The EigenValues of the three components are as follows:

<table>
<thead>
<tr>
<th>Mean</th>
<th>StdDev</th>
<th>Alpha</th>
<th>Factor Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Usefulness (PU)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PU1</td>
<td>5.19</td>
<td>1.23</td>
<td>0.88</td>
</tr>
<tr>
<td>PU2</td>
<td>4.92</td>
<td>1.33</td>
<td>0.73</td>
</tr>
<tr>
<td>PU3</td>
<td>5.21</td>
<td>1.25</td>
<td>0.89</td>
</tr>
<tr>
<td>PU4</td>
<td>5.19</td>
<td>1.30</td>
<td>0.83</td>
</tr>
<tr>
<td>PU5</td>
<td>5.37</td>
<td>1.25</td>
<td>0.66</td>
</tr>
<tr>
<td>Perceived Ease of Use (PEOU)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PEOU1</td>
<td>4.72</td>
<td>1.33</td>
<td>0.86</td>
</tr>
<tr>
<td>PEOU2</td>
<td>4.60</td>
<td>1.29</td>
<td>0.85</td>
</tr>
<tr>
<td>PEOU3</td>
<td>4.99</td>
<td>1.15</td>
<td>0.83</td>
</tr>
<tr>
<td>PEOU4</td>
<td>4.75</td>
<td>1.22</td>
<td>0.78</td>
</tr>
<tr>
<td>PEOU5</td>
<td>5.04</td>
<td>1.51</td>
<td>0.57</td>
</tr>
<tr>
<td>Intention of Use (ITU)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITU1</td>
<td>5.79</td>
<td>1.17</td>
<td>0.85</td>
</tr>
<tr>
<td>ITU2</td>
<td>5.68</td>
<td>1.19</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 2. Summary of descriptive analysis
extracted were 5.436, 1.879, and 1.349 respectively. The percentages of variance explained by the components ranged from 11.245 to 45.302, with a total variance explained of 72.21%.

One-way ANOVA was employed to determine the mean differences between the different gender groups on the major variables (teaching experience, computer training, PU, PEOU, ITU, and Usage). No significant differences were found between gender groups for each variable.

**LISREL Models**

LISREL was a software product designed to estimate and test statistical models of linear relationships among latent and manifest variables. It was an extremely powerful structural equation modeling technique that had been used extensively in research (e.g., Hu et al., 1999). LISREL was then used to analyze the survey data and to perform the analysis towards model testing. Table 3 showed the resulting model testing findings.

For the overall model, the data supported most of the individual causal paths postulated by TAM. Perceived usefulness had a significant direct positive effect on pre-service teacher’s intention to computer use, with standard path coefficient 0.43 (p<0.001). Literally, this coefficient suggested that every unit increment in perceived usefulness would strengthen an individual’s (positive) intention to computer use by 0.43 units. Perceived usefulness also had a direct and significant effect on usage, with standard path coefficient 0.38 (p<0.001). Effect of intention to computer use towards self-reported usage was significant and shown a 0.33 path coefficient. Thus, perceived usefulness had a direct effect, as well as an indirect effect, through the mediating intention to computer use, on usage. Perceived ease of use had positive effects on both perceived usefulness (path coefficient = 0.58) and intention to computer use (path coefficient = 0.15). Although it was a significant direct effect on perceived usefulness (t = 7.41, p<0.000), its effect on intention to computer use was statistically non-significant.

From the R square values, it showed that perceived ease of use explain 33% of the variance in perceived usefulness, while perceived ease of use

**Table 3. Summary of causal path analysis**

<table>
<thead>
<tr>
<th>Construct</th>
<th>Causal Path</th>
<th>Overall</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEOU</td>
<td>PEOU → PU</td>
<td>***0.58</td>
<td>***0.85</td>
<td>***0.47</td>
</tr>
<tr>
<td></td>
<td>PEOU → ITU</td>
<td>0.15</td>
<td>0.07</td>
<td>*0.23</td>
</tr>
<tr>
<td>PU</td>
<td>PU → ITU</td>
<td>***0.43</td>
<td>0.33</td>
<td>***0.43</td>
</tr>
<tr>
<td></td>
<td>PU → USAGE</td>
<td>***0.38</td>
<td>*0.31</td>
<td>***0.45</td>
</tr>
<tr>
<td>ITU</td>
<td>ITU → USAGE</td>
<td>***0.33</td>
<td>***0.61</td>
<td>*0.23</td>
</tr>
<tr>
<td>R²</td>
<td>PU</td>
<td>(0.33)</td>
<td>(0.72)</td>
<td>(0.22)</td>
</tr>
<tr>
<td></td>
<td>ITU</td>
<td>(0.29)</td>
<td>(0.15)</td>
<td>(0.33)</td>
</tr>
<tr>
<td></td>
<td>USAGE</td>
<td>(0.38)</td>
<td>(0.61)</td>
<td>(0.36)</td>
</tr>
</tbody>
</table>

Note: * p<0.05; ** p<0.01; *** p<0.001
and perceived usefulness together explained 29% of the variance in intention to computer use. Intention to computer use and perceived usefulness accounted for 38% of the variance in usage.

This finding was consistent with prior research (e.g., Venkatesh & Davis, 2000) that while “the effect of usefulness on usage was significant,” “the effect of ease of use on usage, controlling for usefulness, was non-significant” with the reason that “ease of use operates through usefulness” (Davis, 1989, pp. 331-332).

The LISREL model was then analyzed on male as well as female data in order to examine the gender differences of the effect of each construct to the model. All data segments provided an overall fit of the model postulated by TAM. However, compared to women, men placed a greater emphasis on perceived ease of use in determining perceived usefulness (0.85 for male; 0.47 for female). On the other hand, women weighted perceived ease of use more strongly in determining intention to computer use than men did (0.07 for male; 0.23 for female). The causal path from perceived usefulness to intention to computer use was non-significant for men. For the case of women, the strong direct significant effect of perceived usefulness in determining intention to computer use showed that was hindered by the overall model, which was not consistent to prior research findings of Venkatesh and Morris (2000).

In agreement with what most literature postulates, the two independent variables, perceived usefulness and perceived ease of use, contributed significantly to the behavioral intention to computer acceptance and actual self-reported usage, accounted for 38%, 61% and 36% of the overall, male and female model respectively.

**MAIN THRUST AND FUTURE TRENDS**

**Key Findings**

The research questions of this empirical study were:

1. What were the emergent constructs that drive the intention and usage behavior of computer use?
2. Did users’ beliefs regarding IT/IS usage differ among genders?
3. To what extent did these effects differ?

The empirical findings were that female and male users’ differ in beliefs, intention, and usage. From the testing summary of the models, the beliefs were influenced and differed among different gender. Female users were influenced by both perceived usefulness and perceived ease of use to their intention and usage of computer, in a more balanced manner. Both factors were significant in predicting intention and usage, though perceived usefulness was stronger in effect ($\beta=0.47$ at $p<0.001$ versus $\beta=0.23$ at $p<0.05$). On the other hand, male users were nearly totally influenced by perceived usefulness to their intention and usage of computer. Perceived usefulness was significant and strong ($\beta=0.85$ at $p<0.001$) while perceived ease of use was non-significant and had only an indirect effect through perceived usefulness toward intention and usage of computer.

**Limitations of the Study**

It was believed that the study on a longitudinal perspective would be conducted in order to gain better understanding about the users’ acceptance behavior. Moreover, this study collected pre-service teachers’ view on “computer” acceptance in a general term. This might also limit us from knowing the differences among other technologies or software applications. Therefore, further studies...
on the area of the acceptance towards different technologies and the acceptance differences on a continuous basis would also be recommended.

**Contributions to Practice**

This empirical study had unique contributions for IT/IS practitioners. Successful system implementation required users to effectively use the system while effective management required a better understanding of the beliefs of users toward IT/IS systems. The findings of gender differences in beliefs and their corresponding beliefs’ strengths could assist in the proactive implementation planning of IT/IS (e.g., user training, system support, etc.) for minimizing the impacts while using IT/IS.

**Contributions to Research**

This study attempted to explore the teachers’ computer acceptance and re-confirmed that perceived ease of use and perceived usefulness were the two independent variables towards computer use. It also revealed the gender differences in the application of the TAM. These findings were definitely important to the design of teachers’ professional development. Viewing teacher training as a kind of remedy for teachers’ inadequacy, teachers’ computer training was, still in many cases, unified and one-off. Without a continuous development plan to teachers (e.g., Bradley, 1991), the problem of acceptance would still be an important barrier to the successful use of computers in education. The study had collected the views of pre-service teachers at a given point of time, however, studies found that the factors to pre-adoption and post-adoption might be different (e.g., Bhattacherjee, 2001; Karahanna, Straub, & Chervany, 1999), that was, a factor contributes positively to acceptance might not necessarily contribute to the same extent and degree after adoption. Sometimes, on the contrary, a factor might hinder further computer use.

**CONCLUSION**

In summary, this study extended the applicability of the technology acceptance model over teachers’ computer use. It validated the instrument and the model framework using survey data from a group of pre-service teachers, as well as reflected gender differences in technology acceptance. We suggested that this model could be applied in the teachers’ computer use context and used to explain the use intention and use behavior. We recommended that these findings to be included in future gender and IT/IS research studies.

**REFERENCES**


Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology:


**KEY TERMS**

Affective Responses to Attitude: They are the responses that reflect evaluations of, and feelings toward the attitude object.

Attitude: An attitude is a disposition to respond favorably or unfavorably to an object, person, institution, or event.

Cognitive Responses to Attitude: They are the responses that reflect perceptions of, and information about, the attitude object.

Computer Attitude Scale: A composite instrument to measure an overall attitude of an user to respond favorably or unfavorably to computer.

Conative Responses to Attitude: They are responses that reflect behavioral inclinations, intentions, commitments, and actions with respect to the attitude object.

Perceived Ease of Use: The degree to which the prospective user expects the target system to be free of effort.

Perceived Usefulness: The prospective user’s subjective probability that using a specific application system will increase his or her job performance within an organizational context.

Technology Acceptance Model (TAM): A model framework which is composed of two fundamental determinants, perceived usefulness and perceived ease of use, to explain computer usage behavior.
APPENDIX: MEASUREMENT ITEMS

<table>
<thead>
<tr>
<th>Perceived Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>PU1</td>
</tr>
<tr>
<td>PU2</td>
</tr>
<tr>
<td>PU3</td>
</tr>
<tr>
<td>PU4</td>
</tr>
<tr>
<td>PU5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Perceived Ease of Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEOU1</td>
</tr>
<tr>
<td>PEOU2</td>
</tr>
<tr>
<td>PEOU3</td>
</tr>
<tr>
<td>PEOU4</td>
</tr>
<tr>
<td>PEOU5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intention to Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITU1</td>
</tr>
<tr>
<td>ITU2</td>
</tr>
</tbody>
</table>

Source: Adopted from Davis, 1989, p. 331
Chapter 7.6
Discourses in Gender and Technology:
Taking a Feminist Gaze

Sheila French
Manchester Metropolitan University, UK

ABSTRACT

The majority of women are not involved in the design, manufacturing or shaping of technology in many Western societies. This is at a time when governments globally see technology as an enabler to economic success. Using feminist scholarship and discourse analysis, this chapter questions why patterns of gender segregation prevail in technology related fields in the United Kingdom. The chapter critically analyses why government policy, and equal opportunities initiatives, have so far largely failed to increase women’s participation. Using examples taken from two educational settings, the chapter uses the narratives of individual’s experiences of technology, their engagement, or lack of engagement with it, to examine the dominant discourses of the field. It is argued that technology discourses, which shape our understanding and identity with technology, are gendered. It is argued that current policies and initiatives, based on giving women equality of access will continue to make little difference. Until gendered dominant discourses of technology are deconstructed and examined; we will not have the tools to address the current situation of gender segregation.

The Connection between masculinity and technology, reflected in women’s under-representation in engineering, and indeed in all scientific and technical institutions, remains strong as we enter a new era of technological change. (Wajcman, 2004)

INTRODUCTION

Globally, governments see new technologies as the enabler of economic success in the global knowledge economy. At the same time the United Kingdom, along with many other Western societies, is experiencing a gender divide in relation to the use, development and design of information and communication technologies (ICT). For sometime it has been recognized that males dominate the use of technologies in all areas of British society (DFEE, 2001; Hellawell, 2001; Wilkinson, 2001) and that gender segregation in ICT occupations persists (EOC, 2004a). Only a few girls are taking
Discourses in Gender and Technology

up computing at an advanced level at school, and universities are experiencing a continued lack of interest in applications by women for computing degree programs (Alexander, 2001b; EOC, 2005). In 1996, 19% of computer science students were reported to be female. Today, there has been little improvement; females account for only 20% of computing graduates in Great Britain (EOC, 2005). In the workplace, women hardly feature in the innovation and production of technology and the computing industry is concerned about the lack of women in the sector. British industry continues to experience major skills shortages of technicians and ICT professionals (DFEE, 2001; EOC 2004b). This is contrary to images in the popular press of women—such as Martha Lane Fox, the co-founder of lastminute.com—who are hailed as heroines of the dot.com industry. In reality men dominate e-commerce start-ups, and there is little involvement of women at the investment level of the industry (Hellawell, 2001). There are signs that women are not involved in the new economy and the new technologies, and “that men are firmly in the driving seat” (Wilkinson, 2001). This has not gone undetected, nor has it been ignored. Over a number of years the lack of women’s participation in science and technology has been addressed in various United Kingdom government policies and initiatives. However, neither the government nor industry has set specific targets in relation to women entering these male dominated industries (EOC, 2004). Gender segregation still prevails and women are still under-represented in the field of technology.

This chapter begins by looking at the emphasis that the United Kingdom government, along with others around the globe, place on the new technologies in relation to the global knowledge economy. The discussion moves on to look at why the current situation of gender segregation is thought to prevail. I present here a critical analysis of government policy and initiatives based on giving equal opportunities to women, most of which have so far largely failed to increase the participation of women. I then introduce feminist theory and discourse analysis to look at “discourses of technology.” Focusing on examples from two educational settings, I use the discussions of individuals’ experiences of technology, their process of engagement, or lack of engagement, with the technology. The aim is to demonstrate that issues of gender and technology are by no means simple. I suggest we should not just focus on giving women equal opportunities to access, training and education in technology, we should instead try to identify and understand more clearly how the dominant discourses around technology come to shape our understanding and identity with technology. It is this I suggest that needs deconstructing before we can address patterns of gender segregation.

ICT AND THE GLOBAL KNOWLEDGE ECONOMY

Technology and innovation feature highly in future economies, and are seen by governments in the United Kingdom and around the globe to be an essential ingredient to becoming internationally competitive (Brooks & Mackinnon, 2001). Training the population in the use of information and communication technologies (ICTs) is seen as a powerful enabler. The lack of access to ICTs does not only lead to exclusion from the new technologies but also to exclusion from the new knowledge economy (Castells, 2000). Training members of society to be computer literate is regarded as essential to participation in the current and future labor market. In government rhetoric about “education” there is a shift in emphasis from being purely concerned with the education of individuals, to a need to ensure the population has the essential skills that will assist with the nation’s wealth creation (Brooks & Mackinnon, 2001; Coffield, 1999). This raises issues about what type of knowledge and skills will be valued by society in the future. It suggests that those
with the knowledge and ability to use the new technologies will be favored for their capacity to contribute to the economy. It suggests that those without the requisite information technology skills will fail to contribute to the economy and, therefore, could be excluded from future prosperity.

UK GOVERNMENT POLICY AND INITIATIVES

The UK government has recognised the gender divide and has proposed a number of initiatives to reverse what they refer to as “the challenge of women’s participation in ICT” (Alexander, 2001b). The aim is to give women access to information technology (IT) in education, the workplace and their social lives. Strategies are being funded to address socially excluded groups, which often include women, to enable them to acquire what are thought to be essential ICT skills for daily life in future economies (DTI, 2004). The aim is to improve the image of IT in education and work; this, it is suggested, will increase women’s participation. Girls will be encouraged to become more enthusiastic about today’s technologies, and it is hoped that they will gain confidence to compete with boys in what the government refers to as the male domination of ICT in the classroom. This is aimed at ensuring that all girls along with other socially excluded groups have the necessary ICT skills to work in and meet the skills demands of the new economy. Female role models will be used to improve the image of IT, to encourage young women to take ICT as a subject at school, and to enter careers related to technology. A change in business attitudes will be promoted in the computing sector to encourage flexible working conditions for parents. This approach is well-meaning, but there are a number of problems with it. Already there is evidence that these policies and initiatives are failing to make little difference.

Firstly this approach treats technology as an artifact that has no political or social values attached to it. This is clearly not the case, as social studies of science and technology have provided strong evidence that technology is not gender-neutral (Adam, 1998; Cockburn, 1985; Wajcman, 1991, 2000). In the home, male members of families still have more access to computers than women do (Richardson & French, 2001). Boys are very often given greater priority of access to computers at home by their parents than girls (Habib & Cornford, 2001; Na, 2001). In education it has been argued that there is a maleness surrounding technology, IT and computing subjects (Woodfield, 2000).

Secondly, there is an assumption that women, if the conditions are right, will want to be involved in the field of technology. However, evidence suggests quite the contrary; women may not want to be involved in computing (Clegg, 2001; Na, 2001). While women are quite able to “do computing,” for many the image of computing and IT is masculine; it is these gendered notions of what “is” technical which lead to girls becoming reticent about taking up computing and technology (Clegg & Trayhurn, 1999; Clegg, Mayfield, & Trayhurn, 1999). Therefore, suggestions that we can use role models to change the image of ICTs lead us to question where the role models will be found if the majority of women in our society remain uninterested in technology. These initiatives presume that given equality of access and the right workplace conditions, women will begin to participate. It suggests that women and men attach the same perceptions and social values to technology—an assumption I believe to be flawed. This approach is misguided in its treatment of technology. Taking this stance is likely to obscure many of the issues related to our social relationship with technology.

In the following discussion feminist theory and discourse analysis will be employed to help us understand the issues involved. We begin by unraveling the existing social and cultural prac-
Discourses in Gender and Technology

practices, which have led to the current situation in the United Kingdom and other Western societies.

TAKING A FEMINIST GAZE

Feminism comprises of one but many different theories and perspectives that have some common understandings and some differences. Feminism is complex, so defining it is controversial. Feminists agree that social and political theory has a history of being “written by men, for men and about men” (Theile, 1986), and that issues of women have been largely ignored or trivialised. Feminists critique any practice where there is an “assumption of male superiority and centrality” in which women’s subordination is taken as a given (Beasley, 1999). However, they do not necessarily agree about how we might bring about changes to any given situation.

Feminists collectively seek to explain women’s oppression and share a belief that women hold an unequal position in society; to use the theories to question the causes for this. The theories do not all suggest the same reasons for oppression or have the same ideals. In their different forms, feminisms collectively “prescribe strategies for women’s liberation” (Tong, 1997). The context in which I will use “discourse analysis” is taken from the critical tradition. Critical feminist work is ultimately political in that it seeks to understand the position of those who suffer most from dominance and inequality. Critical feminists are social critics. They outline their point of view, perspective and their aims, as in this chapter. Often they place their own subjectivity in the research rather than attempting to be neutral observers. The aim is to produce knowledge which might make a positive difference for women. In this case the question being asked is what factors or conditions sustain, legitimise and perhaps condone the current state of social inequality and injustice regarding women’s participation in technology.

DISCOURSE ANALYSIS

Discourse analysis has evolved from a number of theories about how we should study language and text. The way in which I use it here broadly rests on the work of Michel Foucault (1978, 1981). The concept of “discourse” is taken in a quest to understand the relationship between language, social institutions, subjectivity and power. Discourse, as used in a linguistic context, is taken as a system of representation connected to writing or speech. In Foucault’s terms this is not just an analysis of “text” and the spoken word; it is also about how discourses, in his terms, create knowledge or meaning in our social world. Language, in his view, is not necessarily unique to the individual but is shaped by a range of social, political and economic practices. In a way, it places emphasis not only on what one says but also on what one does. Foucault, taking a constructionist theory of meaning and representation, argued that it is “discourse” that gives us meaning, which in turn creates knowledge values or norms in a particular field. There is a range of discourses in society, often overlapping, some of which are more dominant than others. They do not all carry equal weight or power. Foucault argues that the most dominant give meaning to the world and organise social institutions and processes. A dominant discourse in a historical period can then come to be constituted as the norm. Those who hold these beliefs, he argues, have a vested interest in keeping the status quo. This does not mean that individuals cannot contest the discourse, but it may mean that they are perhaps marginalised by what is considered to be the norm in that field. Meaning, therefore, depends on a person’s subjectivity. Subjectivity refers to the conscious and unconscious thoughts and emotions of an individual (Weedon, 1997). In these terms, what we say and what we do are shaped by the discourses we inhabit and the norms and values associated with them. Experiences in the home and at school, such as expectations of the way girls and boys should behave as female
Discourses in Gender and Technology

or male, shape subjectivity. Therefore, if we look at the issues of women’s lack of participation in technology in terms of discourse, it is not only about issues relating to “technology,” “education,” or “careers.” It is also about the context and historical moment in which the discourse resides. In this case the context is also to do with the global knowledge economy and our relationship with technology at this time and the meanings such discourses produce.

To illustrate the complexities and put this work in a historical context I want to return briefly to government policy and initiatives to give an example of such discourses. The following demonstrates how knowledge and power are at work in discourses and how they come to make social meaning.

Consider the following statement made by the UK government, “ensure everyone has the requisite skills for the knowledge economy” (DFEE, 2001). As we read this statement we can see it is about giving the population opportunities to contribute in the workplace. However, it is also linked to the “knowledge economy.” As an educationalist reading this, I would suggest the education of individuals seems to have shifted from pure interest in an individual’s education to securing “economic wealth” for the nation (Brooks & Mackinnon, 2001). Therefore, if we look at the statement in terms of “discourse,” we could question if the statement is perhaps located in a discourse of what we might call “future economic wealth,” or in what Lucey, Melody and Walkerdine (2003) refer to as the discourse of “social capital.” The British government’s political position influences the meaning of this statement, as does my interpretation as an educationalist and feminist. As this discourse of “social capital” is constructed further and propagated by the government in policy and in the media, it becomes a dominant discourse in British society and has the power to shape our norms and values around the new economy.

My argument is that these discourses shape our beliefs and what becomes our “knowledge.” To understand women’s lack of participation, we need to look at norms and values associated with the field of computing and information technology before we can hope to make any difference to women’s lack of participation.

DISCOURSES OF TECHNOLOGY AND GENDER

It is important to define the context in which the terms “computing,” “IT,” and “ICT” are used in this discussion. Computing is a number of disparate and complex practices and technologies where the terms “computing,” “IT,” and “ICT” are used interchangeably. The term “ICT” is relatively new; my interpretation is that it refers more to the “user end” of the technology. The UK government appears to hold a similar interpretation, as the focus is on giving society the necessary “end user” skills for a future economy. In the following discussion, I will use the term “computing” which includes all of these definitions.

Clegg, Mayfield, and Trayhurn (1999) suggest there is not only one “form,” or way, of “doing computing,” and that, this being the case, there is likely to be more than one discourse of computing. They have identified two major discourses; what we can term the “hard end” discourse, dependent on formal methods and mathematical models, and the “soft” or “user end” discourse, in which technology assists or supports organisational systems. Therefore, the meaning of the term “technical” can be different, depending on the context, or frame of reference.

These discourses are reflected in the way we organise “computing,” for example, in education. A computing department in a UK university might place the emphasis of its computing courses on a “mathematical” model of computing (the hard end), whereas a department in a faculty of humanities or business may place more emphasis on
Discourses in Gender and Technology

the human and social dimensions and on the real world uses of computers (the user end). In many academic institutions there is a discourse that suggests that “to do” computing requires superior mental powers linked to those of the mathematician or scientist (Edwards, 1990). In academia, the more traditional view of formal mathematical methods can also be strongly associated with math and “technical machismo” (Mahony & Van Toen, 1990, p. 321). In the disciplines of math and science, there is also a long history of a masculine culture (Hughes, 2001). It has been argued that this association of mathematics and science with masculine culture has turned women away from the subjects. This association suggests a gendered discourse of computing in which women may not wish to participate because it conflicts with their gender identity (Clegg & Trayhurn, 1999). Whilst the field of computing is still defined in this way, it may discourage some, but it cannot solely be responsible for the lack of women’s participation, as women have increased their participation in other seemingly impenetrable “macho” occupations such as medicine. It is also important to note at this juncture that not “all” females are put off by male-dominated occupations; just as not “all” males favour male-dominated occupations only (Hughes, 2001). However, if we look at our lives through the dominant discourses that help us define our identities, we can see that those discourses discussed so far could maintain the status quo in the field of technology.

Through the following two examples, taken from two different case studies carried out in educational settings, I will demonstrate the pervasiveness of these discourses. Both examples are used here to uncover and illustrate the dominant discourses that operate, or serve to maintain how we perceive computing. Both of the studies used are small and are not, therefore, being used to “claim” anything. They are being used to demonstrate the power of discourse in determining what becomes the knowledge, or norm, and values, and, therefore, maintains the current position. The narratives that follow draw on free association techniques (Hollway & Jefferson, 2000), which means that the interviews were unstructured, allowing the individuals to tell their story. The aim was to take a reflective look at what had informed their experiences in relation to their gender identities and their relationship with technology. Using theories of the “gestalt” (Hollway & Jefferson, 2000), I look to the sum of the “whole,” rather than the individual “part”; consequently, the work does not just focus on the subjects’ relationship with technology but also on the influence of the other aspects of their lives and experiences. The students may have constructed their accounts to make sense of the world they inhabit through their subjectivity. I add my interpretation, which I acknowledge is influenced by my own subjectivity and experience (Walkerdine, 1997; Hollway & Jefferson, 2000).

For the first example I use excerpts from the narratives of four university students, two males and two females (aged 18-24) who were studying for a degree in information technology at a British university; this study was carried out in 2003.4

THE NARRATIVES

First we will discuss Simon and Paul. In the following excerpt we can see evidence that for both of them their relationship with technology is embedded in their masculine identities and in several other overlapping discourses around how they played as children and as young adult males. This is not surprising, as there is evidence that the relationship between technology and gender begins in the home where males dominate many of the technologies including computers. When computers first entered the home they were targeted at boys and male hobbyists (Kirkup & Abbott, 1997; Wajcman, 1991). Simon and Paul were not exceptions to this targeting. They had been introduced to computers early in their lives and both of them still played computer games.
Discourses in Gender and Technology

Simon was proud of what he defined as his “technical skills” and how he had recently networked the computers in his home. Simon talked about computers as “something I’m serious at” and explained how, through “trial and error,” he had learnt to build computers.

For Paul the computer is an “object” of some significance in his life. He described how he was comfortable in an environment where there was a computer turned on and how he would leave it on in his bedroom whilst he was working, listening to music or doing other things—“You can click on it and it runs itself... I leave it on in the room and just go back to it.” During the narrative Paul appeared to “confess” to using computers in his leisure time—“It might sound geeky, but yeah, my friends are geeky... I sound like I am always playing games and stuff, but it’s only when I’m sitting in my room and stuff.” Paul appeared to know about the popular images of computing and did not seem to want to be associated with them.

Both Paul and Simon claimed expertise in computing; they said they were “confident,” “good at” computing. Like boys in other studies (Beynon, 1993), they claimed expertise in using computers. I suggest that in their narratives they place themselves in what I have described as the “hard” discourse of computing. For both Simon and Paul, using computers allows them to inhabit a discourse in which they are comfortable. This discourse is located, amongst other issues, in how they played as children, and the way they currently socialise with other males. This promotes personal confidence and a sense of ability with technology, which I suggest is worth preserving. Their interest, as they described it, is fundamental to their identity as males and not just an external interest or passing hobby.

As discussed earlier, boys still have more access to home computing than girls, and there is evidence that parents favour boys over girls in issues of access to computers. It is believed that it is this that has led to girls having less experience of computers than boys before they reach school.

Both Asiya and Karima, the other two students in the study, who are female, both located themselves at the “user end” or “soft end” of any computing discourse. They were very explicit about the “usefulness” of computers in the workplace; neither of them used computers in their leisure time. Both of them expressed how their interest in the computer was gained in their experience at school or at work. “With computers I don’t like the hardware, the technical side of it as much. I understand I need to know about it. I find the technical side difficult, it doesn’t interest me. I would rather be sitting at a PC designing a database than looking inside a PC” (Karima).

Their choice, as with the males, was based around feeling they were good at something. Asiya had found she was “good at using computers” in her secretarial position; she had found time in her position as a secretary to “open program files and discover the PC.” They used the same sort of terms as Simon and Paul—something they were very “interested in,” “good at”—and in Karima’s case she had studied IT at her secondary school and was very enthusiastic—“I absolutely loved it, loved it” (meaning as a subject at school).

Neither of them continued to play computer games or appeared to see much use for the computer outside their university work or in the workplace. They had both developed an interest in computers in their teens, later than Simon and Paul. Both females identified a “use” for computers, but did not perceive that using computers was part of their social activities. This is very much in contrast to Simon and Paul. Simon linked his interest in what he referred to as “building computers” directly to how he played as a boy. We can see in the following how Simon’s “interest” in computers is linked to his identity as a child—“It (computers) are generally like Lego... then it’s like slotting everything in... then it’s setting it up.”

This demonstrates how the gendered attitudes of how Simon played as a child and also gen-
Discourses in Gender and Technology

...dered preferences for toys may have influenced his computer usage. Mead and Piaget have said that children develop a sense of “self” through their play and games (Crossley, 2000). Studies have found that girls and boys play differently and this has implications on the development of their self-identity as girls and boys (Gilligan, 1982). Children’s choices of toys are different. Males are associated with mechanical toys and construction toys such as Lego, which encourage technical confidence (Wajcman, 1991). Girls play with toys such as dolls and stuffed animals (Rheingold & Cook, 1975). All four students were introduced to computers through play, though the females lost interest as they matured. We know that most computer games are targeted at young males and it is likely that most computer games that are available are more popular with boys than girls (Griffiths, 1997). Karima explained she had “grown out of it” and Asiya stated, “I don’t really play, not in the last few years.” The females in this study were no longer motivated to use a computer to play games. Thus they rejected the computer as a toy, changing their view of it in adulthood to that of a useful tool. Computing is something both these female students have an interest in, but I suggest it is not strongly linked to their gendered identities.

What I have shown in this small example is that discourses around technology can extend into the home and other parts of our social lives, including in this case our gender identities as young children and young adults. How we play as children and what are acceptable interests for boys and girls in the home contribute to our gendered identities. All four students have been influenced by childhood and teenage experiences of using computers in the home. I have suggested that both of the male students located themselves in the “hard” end of computing, which in their case is intrinsically linked to them as males and bound up in their masculine identities. Both females were very interested in computers but located themselves in what could be described as the “soft end” or “user” discourse in computing. They could clearly see a use for technology in their lives, but the computer has little to do with their gender identity. The males and females in this study inhabit different discourses; we can identify these are open to interpretation and do not have clear borders.

In the next example, I want to demonstrate how these discourses are perpetuated in another area of education and how they can then influence decision-making of young women when it comes to choosing subjects, courses and careers. We know that educational experiences have the power to influence our perceptions of who we are. Research has identified that “education has a mammoth part to play in gendering social worlds, not only through what is taught but also through how it is,” and therefore perpetuates a society’s norms and values (Evans, 1994, p. 52). There is a great deal of research in the field of gender and education which expands on these issues. In schools and colleges computing and the skills associated with it are still perceived in gendered terms. There are good reasons for this. Early school computers were usually bought and controlled by male teachers in the math and science subject areas. This perpetuated the “hard” discourse and further marginalised girls as the boys showed more interest. A reluctance of girls to embrace the computer was seen early on as a problem of girls’ confidence, rather than rooted within the way technology is perceived (Clegg, 2001). Today there are still only a few women lecturers and teachers in computing and IT at all levels of education in the UK, which may reinforce the idea that it is an area of study most suitable for males. This of course raises the question as to where the government will find their role models.

The following abstracts are taken from a study of 16-18 year old female students studying in two colleges of further education in an inner city area in Manchester, UK. This second study took place in 2004 (French & Saxon, 2006) and was initiated...
Discourses in Gender and Technology

by a number of colleges. They were concerned about the lack of participation of women in their technology courses and the lack of progression of the young women to university and technology industries. The following demonstrates with a few examples how the discourses we have already discussed manifest themselves in this educational setting. Twenty-five female students between the ages of 16 and 18 were interviewed. The narratives demonstrate how the discourses around technology in this study influenced the students’ and teachers’ perceptions of what is a suitable subject for girls to study.

Even before joining their course these female students encountered a gendered discourse around the information technology courses they were interested in—“I couldn’t believe that people said IT was a lad’s (boy’s) course... There were a lot of girls being interviewed at the same time as me, but they didn’t come on the course” (Female, 16 years). She was referring to the course tutor’s remarks during her interview for the course. In light of the description of the course as a course for boys, it is not surprising that so few young women enrolled.

The girls were asked why they thought girls did not enroll in technology courses. There were many duplications of this type of remark; I include only three examples here.

“I think girls prefer hairdressing and girlie type courses.”

“Maybe they think it’s a boy’s thing.”

“A lot of girls think it’s too technical and that they wouldn’t be able to do it.”

We can see a gendered discourse prevails not only through the way the course tutor describes the courses but also around some girls’ perceptions of information technology courses. The EOC recently reported (Fuller, Beck, & Unwin, 2005) that there are still strong gendered perceptions of what are suitable careers for “boys” and “girls” in the UK. Girls are still over-represented in childcare fields, while males are over-represented in subjects related to engineering, motor vehicles and construction. This is despite the fact that as early as 1984, Women into Science and Engineering (WISE) (Henwood, 1996) was launched to encourage women to enter these careers. WISE influenced policy and practice as many small and large-scale initiatives were designed to encourage females to join science, engineering and technology courses. Henwood (1998) has criticised the WISE project for taking an equal opportunities approach based on women’s choice.

In this study, none of the 16-year-old female students said that they were interested in pursuing a career related to technology. Only the female students whose ages ranged from 17-18 years old stated that they were intending to pursue a career associated with their course and technology. There were only 11 females in the 17-18 year age group, out of a total cohort of 83 students; seven were interviewed; all of them were intending to take up a career in computing. It is unlikely that the choices of the younger students had anything to do with their academic achievement. In the UK girls have shown they have ability in the subject area. In 2001-2002 girls achieved 62% of passes at grade C or above in General Certificate of Education – Advanced level (GCE A level) computer studies, whereas the achievement rate for boys was 56% (EOC, 2003). Despite their success in the subject, young women continue to shun technology subjects at the higher levels. In 2004-2005 around half the number of total students in England who took GCE A level in ICT were female. There were 4,510 females compared to 8,370 males, which is an improvement on 2003-2004. However, the numbers of students taking computer studies are a stark reminder of the gender division in this subject area; a total of 5,336 males, but only 493 females took the GCE A level subject (DFES, 2005). It has been shown that there is a direct correlation between
the number of women entering IT or computing degrees at university and those taking the subjects at school and college (Symonds, 2001). In this study, the courses these 17-18 year olds attend are also “feeders” into university degree courses. Whether they can “do” technology is not the issue, they just are not choosing it.

Female tutors and course leaders were in the minority in the two colleges who participated in this study. Several students commented about the male tutors—“The male programming tutor is a bit strange” (15 year old). Most of their personal tutors were male, and another student said “I would rather have a woman to talk to” (15 year old).

Working with male students presented “problems,” or “perceived” problems, some of which had consequences on the course choices as this comment identifies—“I could have gone straight onto the advanced course, but there were no other girls on the course that year. So I did an extra year on my previous course so that I would have some female company” (17 year old). This student’s perceptions had influenced career choice and academic progression. Young men were the dominant group on the IT courses she had studied. Most of the girls interviewed reported they did not like working with the boys. Other students were initially put off by boys, but got used to working with them. Some, as in that quote, had asked to stay in groups with other girls on the course. Having another female in the class made a difference—“If I didn’t have Julie in the class, I would have felt a bit intimidated” (16 year old).

One student had noticed that several girls had started the course with her but did not complete the course. It could be that the thought of a class of mostly male students was worse than the actual experience for some of the students. For some it clearly put them off the course; it meant them having to “fit in.” Some had compromised their career choices and, in other cases, it led them to leave the course before completion. Therefore the dominant number of males on the courses, and the gendered attitudes of the tutors, influenced their experiences.

From the narratives of the university students and the college students we can identify a dominant gendered discourse in computing which is linked to the use of technology in the home, as well as to education. In the previous examples we can see that this discourse is linked to gendered notions of how we view and experience technology. It is this I would argue, which needs deconstructing. I suggest that gendered perceptions of technology started early in life are further perpetuated as women mature. This discourse is pervasive, as it is not just isolated to the field of computing but extends into many other social spheres.

Clearly, if we want to change the current situation, we need to challenge this gendered discourse around technology. This is not easy because it is so pervasive. However, it is possible. We need to look very closely at our gendered identities and our relationship with technology and how, as parents, educators and government, we perpetuate these images. We need to challenge the current values and norms regarding technologies in our society.

**CONCLUSION**

I have shown evidence of a gendered digital divide with regard to technology in British society. I have discussed and criticised government policy. Throughout this discussion, feminism and discourse analysis have been used to unravel the issues and complexities concerning the lack of women’s participation in technology. Through the narratives of these two small studies I have identified gendered discourses around computing and technology. These gendered perceptions relate to our identity and our experiences with technology. These two case studies have shown how this discourse manifests itself in the home and is further perpetuated through our education.
and leads to choices in education and careers. I suggest we need to study this further to enable us to deconstruct this discourse before there is any possibility for change. Gender segregation in ICT is not just an issue of equality; it is also about tackling gendered attitudes and identities in relation to technology. Clearly, it is this we need to address before we can challenge gender segregation and gain women's participation in the future development and use of technology.

ACKNOWLEDGMENT

I would like to thank Diane Saxon, fellow author and research colleague, for her permission to use some of our co-authored work in this chapter. Finally, I would like to thank Helen Richardson for raising my political awareness in our early work together about government policy in relation to gender and technology issues.

REFERENCES


Discourses in Gender and Technology


**ENDNOTES**

1 For a critical discussion of the knowledge economy, see Peters (2001), who argues that the characteristics of the knowledge economy are not necessarily new and that they have been adopted without critical appraisal.

2 Wajcman (2000) gives a full history and analysis of gender and technology studies.

3 Potter (2004) offers a guide to the origins and comparison of discourse analysis dependent on discipline.

4 For a full discussion on this case study, see French (2003).

5 The *Journal of Gender and Education* (Carfax) is a good source of reports of research in this field.
INTRODUCTION

A fundamental consideration when attempting to understand the complex factors leading to the underrepresentation of women in IT is the choice and use of theory. Theories about women and their relationships to information technology and the IT profession guide the conceptualization of the research problem, the methods of data collection, the basis for analysis, and the conclusions that are drawn. However, a criticism of gender and IT research is that the topic of gender and IT is currently undertheorized (Adam, Howcroft, & Richardson, 2001, 2004).

This undertheorization takes on several different forms. First, there are cases in which there is no theory in evidence to guide the conceptualization of the research project or to inform the data collection and analysis. Rather, the focus is typically on compiling and representing statistical data regarding the differences between men and women with respect to technology adoption, use or involvement in the IT profession. This form of undertheorization can be labeled pre-theoretical research. Second, other research, while not explicitly articulating a particular theory, nevertheless, is guided by a theory-in-use. For example, quite often a theory of inherent differences between males’ and females’ relationships to IT is used implicitly to guide data collection and analysis. This form of undertheorization can be labeled implicit-theoretical research. This approach is considered to be a type of undertheorization in that the lack of explicit discussion of a theory makes it difficult for others to discuss, challenge or extend the research. Finally, the body of research that reflects explicit theory-in-use has been shown to have gaps in the theoretical landscape (Trauth, 2002). That is, an argument has been made that current theories about gender and IT do not fully account for the variation in men’s and women’s relationships to information technology and the IT field. This form of undertheorization can be labeled insufficient-theoretical research. It is this third condition that is addressed in this article: the need for new theoretical insights to guide our effort to understand the underrepresentation of women in the IT profession.
Two dominant theoretical viewpoints are currently reflected in the majority of literature about gender and IT: essentialism and social construction (Trauth, 2002). Essentialism is the assertion of fixed, unified and opposed female and male natures (Wajcman, 1991, p. 9). The existence of biological difference between the sexes has led to a tendency to assume that other observed differences between men and women are due to biological determinates as well (Marini, 1990). When applied to the topic of gender and IT, the essentialist theory presumes the existence of relevant inherent differences between women and men with respect to information technology. It uses the observed differences in the participation of women and men in the IT field as evidence of this view. Thus, the causes of gender underrepresentation in IT are attributed to biology. It turns to observed differences in men’s and women’s behavior for explanations of what are believed to be inherent, fixed, group-level differences that are based upon bio-psychological characteristics.

Essentialism underlies research on gender and IT that views gender as a fixed variable that is manipulated within a positivist epistemology (e.g., Dennis, Kiney, & Hung, 1999; Gefen & Straub, 1997; Venkatesh & Morris, 2000). Adam et al.’s (2001) analysis of this perspective points out that focusing on a background literature of psychology, alone, places too much emphasis on individual gender characteristics where a form of essentialism may creep in. Looking only to psychological explanations of observations without giving attention to the influence of context results in a determinist stance with respect to gender.

One inference that could be drawn from an essentialist approach to gender and IT research is that women and men should be treated differently. For example, Venkatesh and Morris (2000) recommend that trainers adopt different approaches toward men and women and that marketers design different marketing campaigns for men and women. Trauth’s critique of essentialist approaches to gender and IT research suggested that one logical extrapolation from this line of thinking to IT workforce considerations would be the creation of two different workforces: a “women in IT” workforce and a “men in IT” workforce. Thus, policies for addressing the gender imbalance would focus on differences between women and men and the equality issue would focus on “separate but equal,” something that was rejected in the arena of racial equality decades ago (Trauth, 2002; Trauth & Quesenberry, 2005; Trauth, Quesenberry, & Morgan, 2004).

The other dominant theoretical perspective focuses on the social construction of IT as a male domain. According to this theory, there is a fundamental incompatibility between the social construction of female identity and the social construction of information technology and IT work as a male domain. This explanation for women’s relationship to information technology looks to societal rather than biological forces. Thus, the causes of gender underrepresentation can be found in both the IT sector and in the wider society.

The literatures of gender and technology in general (e.g., Cockburn, 1983, 1988; Cockburn & Ormrod, 1993; Wajcman, 1991) and that of gender and information technology, in particular (e.g., Adam et al., 1994; Balka & Smith, 2000; Eriksson, Kitchenham, & Tijdens, 1991; Lovegrove & Segal, 1991; Slyke, Comunale, & Belanger, 2002; Spender, 1995; Star, 1995; Webster, 1996) look to social construction theory (Berger & Luckmann, 1966) rather than biological and psychological theories. According to this view, the social shaping of information technology as “men’s work” places IT careers outside the domain of women.

Recommendations for addressing this situation vary. One school of thought based on a multi-year investigation of female underrepresentation in both academe and the workplace in Australia explores the development of strategies to help women fit in to this male domain (e.g., Nielsen, von Hellens, Greenhill, & Pringle, 1998; Nielsen,
von Hellens, Pringle, & Greenhill, 1999; Nielsen, von Hellens & Wong, 2000; Pringle, Nielsen, von Hellens, Greenhill, & Parfitt, 2000; von Hellens, von Hellens, Nielsen, & Trauth, 2001; Pringle, Nielsen, & Greenhill, 2000). Another school of thought focuses on the need to reconstruct the world of computing to become more of a “female domain.” For example, Webster (1996) focuses on the social shaping of female gender identity and the implication for women’s relationship to workplace technologies. Based on analysis of women as a social group in cyberspace, Spender (1995) predicted an influx of “female values” into the virtual world that would accompany increased female presence.

Wajcman’s (1991) analysis of the social constructivist perspective on gender and technology reveals several issues. For example, there is no universal definition of masculine or feminine behavior; what is considered masculine in one society is considered feminine or gender-neutral in another. Further, while gender differences exist they are manifested differently in different societies. Hence, addressing the gender gap in IT employment based upon an assumed “woman’s perspective” is problematic. This analysis suggests a gap in current theory and motivates the articulation of new theory to help us better understand the underrepresentation of women in the IT field.

**MAIN THRUST OF THE ARTICLE**

The need for an alternative theory to account for the underrepresentation of women in the IT workforce emerges from consideration of the assumptions underlying the two prevailing theories discussed in the previous section. The initial work on the Individual Differences Theory of Gender and IT resulted from an analysis of this theoretical gap and used empirical data from a study of gender and IT in Australia and New Zealand (Trauth, 2002; Trauth, Nielsen, & von Hellens, 2003) to make the case for an alternative theory to occupy the space between essentialist theory and social constructionist theory. Subsequent work has focused on greater articulation of this theory (Trauth & Quesenberry, 2005, 2006; Trauth et al., 2004, 2006) and empirical testing of it (Morgan, Quesenberry, & Trauth, 2004; Quesenberry & Trauth, 2005; Quesenberry, Trauth, & Morgan, 2006; Trauth, Quesenberry & Yeo, 2005).

The Individual Differences Theory of Gender and IT addresses the undertheorization of gender and IT by offering an alternative theory that focuses on individual differences among women as they relate to the characteristics of IT work and the IT workplace. This view finds the causes of gender underrepresentation in the varied individual responses to generalized societal influences. Thus, it represents the middle ground between the essentialist and social constructionist theories. In doing so, it investigates the individual variations across genders as a result of the combination of personal characteristics and environmental influences in order to understand the participation of women in the IT workforce. Hence, the focus is on differences within rather than between genders. The theory also views women and men as individuals who possess different technical talents and inclinations and respond to the social shaping of gender in unique and particular ways that vary across cultures. This individual differences theory takes into account the uniformity of social shaping messages conveyed in a culture. However, it also takes into account the varied influence of individual background and critical life events that result in a range of responses to those messages.

The individual differences theory is comprised of three general constructs that, together, explain women’s decisions to enter and remain in the IT field. The individual identity construct includes both personal demographic items (such as age, race, ethnicity, nationality, socio-economic class, and parenting status) and professional items (e.g., industry, type of IT work, etc.). The individual influence construct includes personal character-
istics (e.g., educational background, personality traits, and abilities) and personal influences (e.g., mentors, role models, experiences with computing, and other significant life experiences). The environmental influence construct includes cultural attitudes and values (e.g., attitudes about IT, about women in IT), geographic data (e.g., about the geographical location of one’s work) and economic and policy data (e.g., about the region/country in which one works). The Individual Differences Theory of Gender and IT posits that, collectively, these constructs account for the differences among men and women in the ways they experience and respond to characteristics of IT work, the IT workplace and societal messages about women and men and IT.

CONCLUSION

It is ironic that coincident with a documented need for a deeper understanding of the gender imbalance in the IT field, there is insufficient attention being paid to theorizing gender and IT. Given this need, greater theorization in gender and IT research can contribute in several ways to a better understanding of women’s relationship to information technology. First, it can lead to more theoretically-informed treatments of gender in IT research. Wajcman (2000) has observed that gender is seldom considered a relevant factor in socio-technical studies of IT in context. Second, much of the published work that does focus on gender places emphasis on data analysis rather than theoretical implications and linking these results to the existing body of gender, and gender and IT literature (Adam et al., 2001). Hence, greater explicit use of theory can strengthen the existing body of gender and IT research. Finally, insufficient attention has been paid to the differences among women rather than between women and men with respect to information technology adoption, use and work. The development of the Individual Differences Theory of Gender and IT is intended to address this need by providing additional theoretical insights to help us to better understand the individual and environmental forces that account for the underrepresentation of women in IT. It accomplishes this by focusing on women as individuals, having distinct personalities, experiencing a range of socio-cultural influences, and thus exhibiting a range of responses to the social construction of IT. This, in turn, can facilitate more nuanced studies of gender that explore the multiple identities of women—for example, race and gender, or sexual orientation and gender, or age and gender—and their relationships to information technology.

One stream of future work will explore the role of epistemology and methodology in conducting gender and IT research using the Individual Differences Theory of Gender and IT. Another stream of research will explore the contribution of organizational factors to the underrepresentation of women in IT by focusing on the articulation of workplace factors that enhance and inhibit women’s participation in IT work and women’s varying responses to them. A third stream of research will apply the individual differences theory of gender and IT to an examination of differences in Internet search behavior across a variety of uses.

ACKNOWLEDGMENT

This article is from a study funded by a National Science Foundation Grant (EIA-0204246).

REFERENCES


**KEY TERMS**

**Essentialism:** The assertion of fixed, unified and opposed female and male natures (Wajcman, 1991, p. 9). This theoretical perspective is used to explain the under representation of women in the information technology field by arguing the existence of “essential differences” between males and females with respect to engagement with information technology.

**Individual Differences Theory of Gender and IT:** A social theory developed by Trauth (Trauth, 2002; Trauth et al., 2004) that focuses on within-group rather than between-group differences to explain differences in male and female relationships with information technology and IT careers. This theory posits that the under representation of women in IT can best be explained by considering individual characteristics and individual influences that result in individual and varied responses to generalized environmental influences on women.

**Social Construction:** A theoretical perspective articulated by Peter Berger and Thomas Luckmann (1967) that focuses on social processes and interactions in the shaping of actors. This theoretical perspective is used to explain the under representation of women in the information technology field by arguing that technology—socially constructed as a masculine domain—is in conflict with socially constructed feminine identity.

**ENDNOTE**

1. See Wilson and Howcroft (2000) for an example of how context enriches the analysis of observed differences in behavior toward IT based upon gender.
Chapter 7.8
Socio-Cognitive Model of Trust

Rino Falcone
Institute for Cognitive Sciences and Technology, National Research Council of Italy, Italy

Cristiano Castelfranchi
Institute for Cognitive Sciences and Technology, National Research Council of Italy, Italy

INTRODUCTION

Humans have learned to cooperate in many ways and in many environments, on different tasks, and for achieving different and several goals. Collaboration and cooperation in their more general sense (and, in particular, negotiation, exchange, help, delegation, adoption, and so on) are important characteristics - or better, the most foundational aspects - of human societies (Tuomela, 1995). In the evolution of cooperative models, a fundamental role has been played by diverse constructs of various kinds (purely interactional, technical-legal, organizational, socio-cognitive, etc.), opportunely introduced (or spontaneously emerged) to support decision making in collaborative situations.

The new scenarios we are destined to meet in the third millennium transfigure the old frame of reference, in that we have to consider new channels and infrastructures (i.e., the Internet), new artificial entities for cooperating with artificial or software agents, and new modalities of interaction (suggested/imposed by both the new channels and the new entities). In fact, it is changing the identification of the potential partners, the perception of the other agents, the space-temporal context in which interaction happen, the nature of the interaction traces, the kind and role of the authorities and guarantees, etc.

For coping with these scenarios, it will be necessary to update the traditional supporting decision-making constructs. This effort will be necessary especially to develop the new cyber-societies in such a way as not to miss some of the important cooperative characteristics that are so relevant in human societies.

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.
BACKGROUND

Trust (Ganbetta, 1990; Luhmann, 1990; Dasgupta, 1990), in the general frame described above, might be considered as a socio-cognitive construct of main importance. In particular, trust building is always more recognized as a key factor for using and developing the new interactional paradigm.

Trust should not be made indistinct with security. The latter can be useful to protect – in the electronic domain - from the intrusiveness of an unknown agent, to guarantee an agent in the identification of its partner, to identify the sender of a message (for example, by verifying the origin of a received message; by verifying that a received message has not been modified in transit; by preventing that an agent who sent a message might be able to deny later that it sent the message [He, Sycara & Su, 2001]). With sophisticated cryptographic techniques, it is possible to give some solution to these security problems.

However, more complex is the issue of trust, that must give us tools for acting in a world that is in principle insecure (that cannot be considered 100% secure), where we have to make the decision to rely on someone in risky situations. (Consider the variety of cases in which it is necessary or useful to interact with agents whose identity, history or relationships are unknown, and/or it is only possible to make uncertain predictions on their future behaviors.)

Trust should not be made indistinct with reputation, too. In fact, communicated reputation (Conte & Paolucci, 2002) is one of the possible sources on which the trustier bases its decision to trust or not.

The more actual and important example of the usefulness of trust building is electronic commerce, but we must also consider other important domains of Multi Agent Systems and Agent Theory such as Agent Modeling, Human-Computer Interaction, Computer Supported Cooperative Work, Mixed Initiative and Adjustable Autonomy, Pervasive and Ubiquitous Computing.

In fact, today many computer applications are open distributed systems (with many autonomous components that are spread throughout a network and interacting with each other). Given the impossibility to rule this kind of system by a centralized control regime (Marsh, 1994), it becomes essential to introduce local tools in order to choose the right partnership and at the same time reduce the uncertainty (deriving from the nature of an open distributed system) associated with that choice.

TRUST IN THE NEW TECHNOLOGICAL SCENARIOS

In fact, various different kinds of trust should be modeled, designed, and implemented:

- Trust in the environment and in the infrastructure (the socio-technical system)
- Trust in personal agents and in mediating agents
- Trust in potential partners
- Trust in sources
- Trust in warrantors and authorities.

Part of these different kinds of trust have a complementary relation with each other, that is, the final trust in a given system/process can be the result of various trust attributions to the different components. An exemplary case is one’s trust in an agent that must achieve a task (and more specifically in its capabilities for realizing that task) as different from one’s trust in the environment (hostile versus friendly) where that agent operates, or again as different from one’s trust in a possible third party (arbitrator, mediator, normative systems, conventions, etc.) able to influence/constrain the trustee and representing a guaranty for the trustier (Castelfranchi & Falcone, 1998; Falcone & Castelfranchi, 2001).

Therefore, the “sufficient” trust value of one single component cannot be established before
Socio-Cognitive Model of Trust

evaluating the value of the other components. In this regard, it is very interesting to characterize the relationships between trust and (partial) control (Castelfranchi & Falcone, 2000).

It is important to underline how trust is in general oriented towards not directly observable properties. It is, in fact, based on the ability to predict these properties and to rely or not to rely on them. Thus, it is quite complex to assess the real trustworthiness of an agent/system/process, not only because - as we have seen - there are many different components that contribute to this trustworthiness, but also because the latter is not directly observable (see [Bacharach & Gambetta, 2001] about signs of trust). The important thing is the perceived trustworthiness that is, in its turn, the result of different modalities of the trustier’s reasoning about direct experience; categorization; inference, and communicated reputation.

SOCIO-COGNITIVE MODEL OF TRUST

The Socio-Cognitive model of trust is based on a portrait of the mental state of trust in cognitive terms (beliefs, goals). This is not a complete account of the psychological dimensions of trust. It represents the most explicit (reason-based) and conscious form. The model does not account for the more implicit forms of trust (for example, trust by default, not based upon explicit evaluations, beliefs, derived from previous experience or other sources) or for the affective dimensions of trust, based not on explicit evaluations but on emotional responses and an intuitive, unconscious appraisal (Thagard, 1998).

The word trust means different things, but they are systematically related with each other. In particular, three crucial concepts have been recognized and distinguished not only in natural language but also in the scientific literature. Trust is at the same time:

- A mere mental attitude (prediction and evaluation) toward another agent, a simple disposition;
- A decision to rely upon the other, i.e., an intention to delegate and to trust, which makes the trustier “vulnerable”(Mayer, Davis, & Schoorman, 1995);
- A behavior, i.e., the intentional act of trusting, and the consequent relation between the trustier and the trustee.

In each of the above concepts, different sets of cognitive ingredients are involved in the trustier’s mind. The model is based on the BDI (Belief-Desire-Intention) approach for modeling mind that is inspired by Bratman’s philosophical model (Bratman, 1987). First of all, in the trust model only an agent endowed with both goals and beliefs can “trust” another agent. Let us consider the trust of an agent $X$ towards another agent $Y$ about the ($Y$’s) behavior/action $\alpha$ relevant for the result (goal) $g$ when:

- $X$ is the (relying) agent, who feels trust; it is a cognitive agent endowed with internal explicit goals and beliefs (the trustier)
- $Y$ is the agent or entity that is trusted (the trustee)
- $X$ trusts $Y$ about $g/\alpha$ and for $g/\alpha$.

In the model $Y$ is not necessarily a cognitive agent (for instance, an agent can - or cannot - trust a chair as far as to sustain his weight when he is seated on it). On the contrary, $X$ must always be a cognitive agent: so, in the case of artificial agents we should be able to simulate these internal explicit goals and beliefs.

For all the three notions of trust defined above (trust disposition, decision to trust, and trusting behavior), we claim that someone trusts some other one only relatively to some goal (here the goal is intended as the general, basic teleonomic notion, any motivational representation in the agent: desires, motives, will, needs, objectives, duties,
Socio-Cognitive Model of Trust

utopias, are kinds of goals). An unconcerned agent does not really “trust”: he just has opinions and forecasts. Second, trust itself consists of beliefs.

Since Y’s action is useful to X (trust disposition), and X has decided to rely on it (decision to trust), this means that X might delegate (act of trusting) some action/goal in his own plan to Y. This is the strict relation between trust disposition, decision to trust, and delegation.

The model includes two main basic beliefs (we are considering the trustee as a cognitive agent, too):

- **Competence Belief**: a sufficient evaluation of Y’s abilities is necessary. X should believe that Y is useful for this goal, that Y can produce/provide the expected result, and that Y can play such a role in X’s plan/action.

- **Willingness Belief**: X should think that Y not only is able and can do that action/task, but Y actually will do what X needs (under given circumstances). This belief makes the trustee’s behavior predictable.

Another important basic belief for trust is:

- **Dependence Belief**: X believes -to trust Y and delegate to it- that either X needs it, X depends on it (strong dependence), or at least that it is better to X to rely on it, rather than not to rely on it (weak dependence). In other terms, when X trusts someone, X is in a strategic situation: X believes that there is interference and that his rewards, the results of his project, depend on the actions of another agent Y.

Obviously, the willingness belief hides a set of other beliefs on the trustee’s reasons and motives for helping. In particular, X believes that Y has some motives for helping it (for adopting its goal), and that these motives will probably prevail -in case of conflict- on other motives, negative for it. Notice that motives inducing adoption are of several different kinds: from friendship to altruism, from morality to fear of sanctions, from exchange to common goal (cooperation), and so on. This is why, for example, it is important to have common culture, shared values, and the same acknowledged authorities between trustier and trustee.

Another important characteristic of the socio-cognitive model of trust is the distinction between trust “in” someone or something that has to act and produce a given performance thanks to its internal characteristics, and the global trust in the global event or process and its result, which is also affected by external factors like opportunities and interferences.

Trust in Y (for example, “social trust” in strict sense) seems to consists in the first two prototypical beliefs/evaluations identified as the basis for reliance: ability/competence (that with cognitive agents includes knowledge and self-confidence), and disposition (that with cognitive agents is based willingness, persistence, engagement, etc.). Evaluation about external opportunities is not really an evaluation about Y (at most the belief about its ability to recognize, exploit and create opportunities is part of our trust in Y). We should also add an evaluation about the probability and consistence of obstacles, adversities, and interferences.

Trust can be said to consist of or better to (either implicitly or explicitly) imply the subjective probability of the successful performance of a given behavior a, and it is on the basis of this subjective perception/evaluation of risk and opportunity that the agent decides to rely or not to rely on Y. However, the probability index is based on, and derives from those beliefs and evaluations. In other terms, the global, final probability of the realization of the goal g, i.e., of the successful performance of a, should be decomposed into the probability of Y performing the action well (internal attribution) and the probability of having the appropriate conditions (external attribution) for the performance and for its success, and of not having interferences and adversities (external
**Socio-Cognitive Model of Trust**

The degree of trust is used to formalize a rational basis for the decision of relying and betting on $Y$. A “quantitative” aspect of another basic ingredient is relevant: the value or importance or utility of the goal ($g$). In sum, the *quantitative dimensions of trust are based on the quantitative dimensions of its cognitive constituents.*

If we call $DoTXY\tau$ the degree of trust of an agent $X$ about $Y$ on the task $\tau=(a,g)$ we have:

$$DoTXY\tau = DoCX[OppY(a,g)] \cdot DoCX[AbilityY(a)] \cdot DoCX[WillDoY(a,g)]$$

Where:

- $DoCX[OppY(a,g)]$, is the degree of credibility of $X$’s beliefs about $Y$’s opportunity of performing $a$ to realize $g$;
- $DoCX[AbilityY(a)]$, the degree of credibility of $X$’s beliefs about $Y$’s ability/competence to perform $a$;
- $DoCX[WillDoY(a,g)]$, the degree of credibility of $X$’s beliefs about $Y$’s actual performance;
- $DoCX[WillDoY(a,g)] = DoCX[IntendY(a,g)] \cdot DoCX[PersistY(a,g)]$
  (Given that $Y$ is a cognitive agent)

In any circumstance, an agent $X$ endowed with a given goal, has three main choices:

i) To try to achieve the goal by itself;
ii) To delegate the achievement of that goal to another agent, $Y$;
iii) To do nothing (relatively to this goal), renouncing.

Considering the simplified scenario in which only (i) and (ii) are the possible choices we have the Figure1.

Where if $U(X)$ is the agent $X$’s utility function, more specifically:

---

**Figure 1.**

```
To do by itself
  \[ U(X) \]
Failure
  \[ U(X) \]

Delegation
  \[ U(X) \]
Success
  \[ U(X) \]
Failure
  \[ U(X) \]
```

**attribution**. This decomposition is important because:

a. The trustier’s decision might be different with the same global probability or risk, depending on its composition (for example, for personality factors);

b. Trust composition (internal vs. external) produces completely different intervention strategies: to manipulate the external variables (circumstances, infrastructures) is completely different from manipulating internal parameters.

The idea that trust is gradable is usual (in common sense, in social sciences, in Artificial Intelligence). However, since no real definition and cognitive characterization of trust is given, the quantification of trust is quite *ad hoc* and arbitrary, and the introduction of this notion or predicate is semantically empty. On the contrary, in the socio-cognitive model of trust there is a strong coherence between the cognitive definition of trust, its mental ingredients, and, on the one side, its value, on the other side, its social functions. More precisely the latter are based on the former.

A degree of trust of $X$ in $Y$ is grounded on the cognitive components of $X$’s mental state of trust. More precisely the *degree of trust is a function of the subjective certainty of the pertinent beliefs.*
Socio-Cognitive Model of Trust

- $U(X)p^+$, the utility of the $X$’s success performance;
- $U(X)p^-$, the utility of the $X$’s failure performance;
- $U(X)d^+$, the utility of a successful delegation (the utility due to the success of the delegated action);
- $U(X)d^-$, the utility of a failure delegation (the damage due to the failure of the delegated action).

In the previous scenario, in order to delegate we must have:

$$DoTXYτ * U(X)d^+ + (1 - DoTXYτ) U(X)d^- >\]
$$DoTXXτ * U(X)p^+ + (1 - DoTXXτ) U(X)p^-$$

where $DoTXXτ$ is the selftrust of $X$ about $τ$.

More precisely, we have:

$$U(X)p^+ = \text{Value}(g) + \text{Cost}[\text{Performance}(X)],$$

$$U(X)p^- = \text{Cost}[\text{Performance}(X)] + \text{Additional Damage for failure}$$

$$U(X)d^+ = \text{Value}(g) + \text{Cost}[\text{Delegation}(X Y)],$$

$$U(X)d^- = \text{Cost}[\text{Delegation}(X Y)] + \text{Additional Damage for failure}$$

Where it is supposed that it is possible to attribute a quantitative value (importance) to goals and where the costs of the actions (delegation and performance) are supposed to be negative.

CONCLUSION

The Socio-Cognitive model of trust analyzes the basic elements on which trust is founded in terms of the cognitive ingredients of the trustier. In fact, the richness of the referred model (trust is based on many different beliefs) allows us to distinguish between internal and external attributions (to the trustee) and for each of these two attributions it allows us to distinguish among several other subcomponents such as: competence, disposition, un-harmfulness, and so on.

The model introduced a degree of trust instead of a simple probability factor since it permits us to evaluate trustfulness in a rational way.

In other words, if we understand what precisely the basic ingredients of trust are, we would be able to better model and build artificial systems in which this attitude should be present.

FUTURE TRENDS

One of the main aspects that should be analyzed in the next few years is the dynamics of trust and the possibility of introducing all the dynamic aspects in the computational setting in which humans and machines will work together. Trust is a dynamic phenomenon in its intrinsic nature (Falcone & Castelfranchi, 2001). Trust changes with experience, with the modification of the different sources it is based on, with the emotional state of the trustier, with the modification of the environment in which the trustee is supposed to perform, and so on. But, trust is also influenced by trust, itself, in the same specific interaction: for example, how trust creates a reciprocal trust; how the fact that $A$ trusts $B$ can actually increase $B$’s trustworthiness; and so on. In other words, in a computational model of trust relationships we have to consider all the dynamical aspects of the trust phenomenon.

REFERENCES


Socio-Cognitive Model of Trust


**KEY TERMS**

**Cyber-societies:** The set of natural, artificial, and virtual agents connected and interacting with each others through natural and artificial infrastructures within virtual institutions.

**Reputation:** the estimated trustworthiness in an agent as derived from the communicated opinions of other parts (directly or indirectly received); the resulting and emergent “common opinion” about the agent’s trustworthiness.

**Task:** An action and/or a goal an agent has to realize as delegated by another agent; thus – in the opposite perspective - the couple action/goal that an agent intentionally delegates to another agent; where at least the delegating agent knows one between the action and the goal.

**Trust:** The attitude of an agent to delegate a part of its own plan/goal to another agent and rely upon it in a risky situation (possible failure) on the basis of its own beliefs about the other agent and on the environment in which it operates.

**Trustee:** the trusted agent in the trust relationship.
**Socio-Cognitive Model of Trust**

**Trustier:** the trusting agent in the trust relationship.

**Ubiquitous Computing:** The trend of the technological development to integrate into any kind of object information processing and communication capabilities.

*This work was previously published in Encyclopedia of Information Science and Technology, Vol. 5, edited by Mehdi Khosrow-Pour, pp. 2534-2538, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
ABSTRACT

This chapter is on cultural aspects of information-communications systems embedded into new media environment and invisible e-technologies, and on a new age of social responsibility for information technology professionals. Besides the key issues in information technology development that create smart environment and ambient intelligence, the chapter also discusses digital e-culture and the new media role in cultural heritage. From the viewpoint of information technology, the current information-communications systems converge with media. This convergence is about tools-services-content triangle. Thus, we are confronted with a new form of media mostly presented with the term digital, reshaping not only media industry but also a cultural milieu of an entire nation on a regional and global basis. The discussion follows on the World Library idea that is rebuilding with new form of World Memory (World Brain), the shift from visible culture domination to the domination of invisible culture in the world of e-technologies predominance. From this scenario, information technology professionals coping with information systems projects, e-services development, and e-content design have more cultural responsibility than in the past when they worked within closer and inner cultural horizons and when their misuse of technologies had no influence on culture as a whole.

INTRODUCTORY REMARKS

The information society is, above all, an economic concept but with important social and cultural implications. The new forms of direct access to information and knowledge create new forms of e-culture. E-culture is a part of a culture. It not
only concerns users but the community of information professionals as well. When one speaks of information technology, it is always from the Western point of view, whereas e-technology (especially its applications) takes place throughout the world, and every culture has a different understanding of it. The shift to an e-culture at the level of society in general is translated to the individual level, enabling cultural change to be described empirically. The term e-culture refers to the diffusion of new technology, its application for various purposes (especially information and communication), and shifts in related attitudes, values, and norms. E-technology may not be gnawing at the roots of our culture, but those roots are gradually absorbing it. As with all innovation, cultural or otherwise, this technology will reinvigorate, transform, and inspire older cultural forms. We are living in the era of globalization, the information economy, with borderless communities and multiple citizenships. E-culture literacy and attainment will require serious attention to new infrastructure, to the building blocks and platforms for e-culture. These are critical issues for the pursuit of information professionals’ excellence, for creativity in an information society, as well as for fundamental imperatives for commerce and trade in a new media environment.

The main notion of the following text is on cultural aspects of information-communications systems embedded into a new media environment, on invisible technologies, and on a new age of cultural responsibility for information technology professionals. The key issues in information technology development that create invisible e-technologies and smart environments are under e-culture influence. From the viewpoint of information technology, the current information-communications systems converge with media. This convergence is about tools-services-content triangle. Thus, we are accepting a new form of media mostly presented with the term digital, reshaping not only media industry but also a cultural milieu of an entire nation on a regional and global basis. The discussion follows on the new e-technology and information-communications systems convergence as the basis for defining pervasive computing and positive e-technologies. The findings at the end of this chapter explain the process of a fundamental cultural shift from the computer-based information technology to the computerless (invisible) e-technologies in which the e-culture is the essential factor of the success. The discussion section is about the role of information technology professionals coping with information systems projects, e-services development, and e-content design. They have more social responsibility than in the past when they worked within closer and inner cultural horizons and when their misuse of technologies had no influence on culture as a whole.

**BACKGROUND ON INFORMATION COMMUNICATIONS SYSTEMS AND NEW MEDIA**

One of the most valuable and essential processes that humanity can engage in and which is, therefore, essential to look at in terms of information technologies, is the process of self-determination. The principal of self-determination of people was embodied as a central purpose of the United Nations in its 1945 charter. The purposes of the United Nations are to develop friendly relations among nations based on respect for the principle of equal rights and self-determination of nations, and to take other appropriate measures to strengthen universal peace. Resolution 1514 (XV) of December 14, 1960, containing the Declaration on the Granting of Independence to Colonial Countries and Peoples, stated that all nations have the right to self-determination; by virtue of that right, they freely determine their political status and freely pursue their economic, social, and cultural development (United Nations, 1960). In the 1990s,
these issues continued to be highly relevant as numerous people around the world strove for the fulfillment of this basic right of self-determination. The UN General Assembly in 1995 again adopted a resolution regarding the universal realization of the right of nations to self-determination. Thus, the General Assembly reaffirmed the importance for the effective guarantee and observance of human rights and of the universal realization of the right of nations to self-determination (United Nations, 1995). By this, we see that self-determination is tied to all aspects of life: political, economic, social, and cultural. It is ultimately about how we choose to live and allow others to live together on this planet. Furthermore, information technology plays a key role in current economic and social affairs, so the information technology specialists/professionals have much more social responsibility than other professions. Information and communication technologies and networking infrastructures are playing an expanding role in supporting the self-determination of people and emergent nations. Access to information and the facilitation of communication provides new and enhanced opportunities for participation in the process of self-determination. It gives the potential to enhance political, economic, social, educational, and cultural advancement beyond the scope of traditional institutions and forms of governance.

The next step in recognizing cultural and social dimensions of information technology on the international scene is regarding the Council of Europe document, Declaration of the Committee of Ministers on human rights and the rule of law in the information society (Council of Europe, 2005). The Declaration recognizes that information and communication technologies are a driving force in building the information society with the convergence of different communication media. It also stressed that building societies should be based on the values of human rights, democracy, rule of law, social cohesion, respect for cultural diversity, and trust between individuals and between nations, and their determination to continue honoring this commitment as their countries enter the Information Age.

Vannevar Bush (1945) predicted that the advanced arithmetical machines of the future would be (a) electrical in nature, (b) far more versatile than accounting machines, (c) readily adapted for a wide variety of operations, (d) controlled by instructions, (e) exceedingly fast in complex computation, and (f) capable of recording results in reusable form. The new computer devices as smart devices, linked through communications systems, are creating new forms of information-communications systems. Thus, the new form of information appliances and ubiquitous information technology creates the basis for the concept of an information-processing utility. Based on interactive and ubiquitous carriers of information, the first generation of new information systems evolved to provide easy communication over time and space barriers. Thus, the new information systems are media. They are virtual communication spaces for communities of agents interested in the exchange of goods and knowledge in a global environment. Further promising technologies are pervasive computing and augmented reality. The vision of pervasive computing is, to some extent, a projection of the future fusion of two phenomena of today: the Internet and mobile telephony. The emergence of large networks of communicating smart devices means that computing no longer will be performed by just traditional computers but rather by all manners of smart devices. From these notions, it is evident that information-communications systems open the way to information society development. The information society is based on the new (digital) media that provides vast opportunities for information/content networking. New organizational networks are built, cutting across national borders and interests. The networks themselves increasingly may take precedence over nation-states as the driving factor.
Information-Communications Systems Convergence Paradigm

in domestic and foreign affairs. At the same time, native communities have been actively engaged in creating and utilizing such networks with increasing participation and sophistication.

We are entering the era of new media. New media are tools that transform our perception of the world and, in turn, render it invisible or visible. Information technology (IT) professionals must understand new problems, considering the role of e-technologies in the integration and interaction between cultures. It is apparently true with tera architecture of the sensor networks that will transform business, healthcare, media, and e-culture itself. A new form of information-communications systems boosts intelligent networks with the majority of computers that are invisible and disposable. The IT professionals have the challenge in turning all that data into useful and meaningful information and in resolving cultural and privacy issues that accompany pervasive networked computing and ambient intelligence. IT professionals are confronted with the stage when e-technologies extract analytic values from social networks turning information issued by sensors and other data sources into knowledge management systems.

Defining new media is hard work. If we begin to use voice or books in an innovative fashion, we have just made old media into new media. Whatever we define as new media now would be old media as soon as we add innovations. We cannot define new media strictly based on the use of new technology for distance communication, since technology is always changing. What is new media today will be passé tomorrow. If we try to define new media by process rather than by structure, we are still in trouble. Whatever we define new media as today no longer will be valid tomorrow as technology changes the structures and processes. Trying to define the limits of change is a futile effort due to the very nature of change. This means that the regulation of new media is also an exercise in utility. The experiences with first-generation media platforms showed that in order to take advantage of the potentials and chances offered by new media, we need to explore their features and learn how to use them effectively and to build them efficiently. In short, we need to develop innovative concepts, frameworks, and methodologies for the design, realization, and management of the new media. The new media offer unprecedented opportunities and potentials for positively changing almost any aspect of our lives. The growing importance of new media and the demand for appropriate platforms have given rise to the development of innovative technologies and components for such media. Consequently, we can now observe the first generation of media platforms and the first management approaches for such platforms.

The evolution in convenient, high-capacity storage of digital information is one of the enabling technologies for new media. Disk drives that allow local storage, retrieval, and manipulation of digital content are increasing in capacity and falling in price. The current TV experience will evolve into a highly personalized process. Consumers have access to content from a wide variety of sources tailored to their needs and personal preferences. New business models and opportunities for the various providers in the value chain will evolve in an organic market focused on addressing individuals directly with new services. This will allow content providers to respond more effectively to audience needs. Digital media and the emerging communication technologies have created an overabundance of programs and information available from which each consumer can choose. The consumer will need new solutions enabling smart and active decision making over viewing preferences, such as a personal filter for the multitude of choices, dynamically adapting to changing needs and preferences.

Communications technology is available for the support of highly complex interenterprise service networks that support new services (Negroponte, 1996). Altogether, this creates a new view on the product, emphasizing the utility of
the package (product and services) instead of the product itself. Analyzing the lifecycle of a product is crucial for synthesizing and specifying new types of benefits for a customer. Therefore, modern manufacturers have to provide benefits to customers. Questions are what could be the benefits and what kind of utility may be beneficial for the customer. Based on that exercise, they have to come up with appropriate business concepts based on new media e-culture.

THE CULTURE AND INVISIBLE CULTURE

To cope with the new culture space in the context of information-communications systems embedded into new media environment, there is a need for basic definitions on culture, new media, and digital e-technologies. The great advances in culture come not when people tried to impose the values of one culture to the exclusion of all others, but rather when modern individuals try to create structures that are more exciting by combining elements from different cultures. The current information technology is capable of recording universal standards and particulars around the world, and it opens new ways and sources for creativity and global cultural heritage. Hence, information technology must reflect the full range of human existence, the values, the culture, and the entire knowledge. At the same time, the new e-culture is born interacting with e-technologies, and it exists in new cultural ecology. This new cultural ecology stimulates the development of a new trio (triple convergence) consisting of e-technologies, e-culture, and e-society.

The culture is a shared set of manifest and latent beliefs and values (Sackmann, 1991). It helps people to categorize and predict their world by teaching them about habits, rules, and expectations from the behaviors of others. Culture also molds the way people think—what their motivations are, how they categorize things, what inference and decision procedures they use, and the basis on which they evaluate themselves. Most of other definitions are too narrow. Sociologists have focused on behaviorist definitions of culture as the ultimate system of social control. In this system, people act appropriately and monitor their own standards and behaviors. Thus, the culture consists of the learned ways of group living and group responses to various stimuli; sociologists describe the content of the culture as the values, attitudes, beliefs, and customs of a society. Media theorists have explored the interplay of culture and technology, which has led to an emphasis on some aspects of culture. The new approaches are considering cultural ecology as consisting of new media in which various types of media are translated into a common digital form that is accessible within a single framework.

Today’s networked media allow each user to participate actively in the creation of cultural expressions, which we perceive simultaneously and with immediate proximity. The new culture is emerging due to the use of digital technology (e-technology). At the same time, there is a strong relation between values promoted by the new digital e-culture and the traditional moral values created by the major world cultures. These notions open the new contextual approaches on culture in the information society. The culture related to the information society is about three contextual elements (Figure 1.)

The forces of globalization and technology development are paradoxical by nature, offering both threats and opportunities for cultural diversity. Yet, the information society is currently perceived only as an economic imperative in a new environment shaped by rapid information technology developments, based on visions shaped primarily by technologist and business concerns and priorities. The prevailing options embedded in these visions, such as globalization based on cultural homogenization, are questionable not only from a political and social standpoint but also in economic terms. Citizens around the world
are becoming increasingly concerned about the way accelerating processes of globalization and technological innovation are leading to cultural homogenization and immense concentrations of financial power. Globalization generally is seen to be a phenomenon driven primarily by economic interests. As such, it has neither moral content nor values. Therefore, it could be independent from culture.

**Cultural Diversity and the Information Society**

Cultural diversity potentially can become a key asset in the information society, despite the fact that the culture could be defined as an obstacle. It is clear from the previous discussion on globalization that the economic forces of globalization pose a serious threat to cultural identity.

---

*Figure 1. Triple context of culture in information society*

*Figure 2. Tools-services-content triangle as the basis for information society*
Information-Communications Systems Convergence Paradigm

Information technologies are not only the tools that accelerate the pace of globalization, but they are also becoming the key means of access to any good or service. One could thus argue that cultural diversity is an obstacle. Cultural diversity is essentially a question of communication, both internally to one’s own culture and externally with distributed cultures. Thus, one of the central issues is not only access to new e-technologies but also the capability of manipulating new media in order to participate actively in communicational exchanges.

Technology literacy is needed to operate IT effectively. Cultural literacy refers to the ability of an individual (or community) to relate to the services made by one’s own cultural heritage and with access to another culture in a positive way. This means learning from both similarities and differences, being able to reject some aspects, and accepting others. Cultural literacy thus lies at the heart of the possibility of communication in a context of cultural diversity. Content is about Creativity. Creativity is the factor lying at the intersection of motivation and competence, and it covers both the individual and the collective levels. Collective creativity is of greater importance if we are aiming for a shift toward an information society in which a given community will depend on the collective creativity of its social and economic individuals and organizations (with accepting collective memory and brain).

Visible Culture

Western culture has had a strong bias toward the so-called fine arts, such as painting and sculpture. These expressions of culture, which are continuously visible, were more significant than the performance arts (theatre, film, music), which are only visible when they are being played. One of the key elements of great visible culture is uniqueness. Thus, the challenge of universal standards has brought the question of uniqueness back to the fore. To communicate internationally, we need global standards that bring the risk of reducing everything to one mode of expression.

The world of telephony offers an interesting case in point. We clearly need standards and uniform rules for telephones, yet every conversation on those telephones still can be different (Veltman, 1997). In this imperative to record the particular as well as the universal, creativity is only one essential element. The major cultures of the world owe much of their greatness to the fact that they have a recorded tradition, which stabilizes the corpus but also ensures the possibility of a cumulative dimension, which is reflected in terms such as cultural heritage. Even so, there are many skills in the craft tradition relating to culture that remain oral and invisible.

Invisible Culture

Many people favor material culture because it is visible and easily recognized. Culture is about more than objects in visible and tangible places. For example, many computer users are accustomed to thinking of computers as tools for answering questions. We need to think of them as tools for helping us to understand which questions can be asked, to learn about contexts when and where questions are not asked, of knowing that there are very different ways of asking the same thing. If software continues to be dominated by one country and if the so-called wizards of those programs all rely on the questions of that single country, then many potential users of computers inevitably will be offended, and it is likely that they will not use the programs.

The base assumption is that culture for us is invisible (invisible culture). As workers do not know that they are participants accepting entire organization values, the culture of an organization is invisible (Cooke & Lafferty, 1989). Yet, it is all-powerful. Therefore, this assumption is important for invisible e-culture. We are working in organizations that actually drive our behavior and performance in a way that most of the time is
Information-Communications Systems Convergence Paradigm

A definition of culture includes the way we do things. A way to look at culture is actually to have some outside expert or outside person come in and question the way things are going, which can begin to give some insight to the people inside the organization about their own culture (Cleary & Packard, 1992). However, the culture is also considering exchange of ideas, thoughts, and beliefs, and it helps people realize that things can be done differently somewhere else. The success in one culture does not mean success in another culture, and there are actually many ways to succeed, change, and live. People realize that they can act on some other level of culture and act on their own culture.

E-Culture

We could define culture as the beliefs, behaviors, languages, and entire way of life of a particular time or group of people. Culture includes customs, ceremonies, art, inventions, technology, and traditions. The term also may have a more specific aesthetic definition and can describe the intellectual and artistic achievements of a society. The new world economy develops in e-culture and characterizes with the fast, open access to information and the ability to communicate directly with nearly anyone anywhere (Kanter, 2001). This sets e-culture apart from traditional environments.

In a first approximation, one could say that an e-culture is emerging from the convergence of communication and computing along with globalization and the penetration of e-technology in the smallest corners of our lives. The advent of information and communication technology goes hand in hand with changes in attitudes, skills, and behaviors that play a central role in daily life. The advent of an e-culture is correlated with terms of a broad definition of culture. This concerns the culture of a society with both invisible and material characteristics. E-technology as a part of the cultural information may be classified as e-invisible culture, but the outputs of that technology (information appliances) may range among the material (visible) cultural products.

The shift to an e-culture at the level of society is translated to the individual level enabling cultural change to be empirically described. The term e-culture refers to the diffusion of new technology; its application for various purposes (especially information and communication); and shifts in related attitudes, values, and norms. The human thinking and behavior are changing gradually by information and communication technology. E-technology may not be gnawing at the roots of our culture, but those roots gradually are absorbing it. As with all innovation, cultural or otherwise, this technology will reinvigorate, transform, and inspire older cultural forms. We are living in the era of globalization, the information economy, with borderless communities and multiple residencies. E-culture literacy and attainment will require serious attention to new infrastructure and to the building blocks and platforms for e-culture. These are critical issues for the pursuit of information professionals’ excellence, for creativity in an information society, as well as for fundamental imperatives for commerce and trade in new media environments.

E-TECHNOLOGIES AND UBQUITOUS INFORMATION (DIGITAL) APPLIANCES

The Internet is without precedent because of two key features: its interactive and communicative natures. It is not a commodity in the sense that you can go out and buy a TV. You cannot go out and buy a net. The key word here is interactivity. Interactivity implies a dialogue of some kind, a changing response based on changing stimuli. There is much talk of interactive Web sites, but even the best of these choose from a preprogrammed set of possibilities in order to give the
illusion of being interactive. You have interactivity over the telephone, and you have interactivity in a face-to-face dialogue. However, you do not have interactivity in traditional analogue television. The prosperity of the net is that it permits simultaneous interactivity with thousands and, perhaps, millions of people worldwide. This is a first in the history of humankind.

With the new and upcoming information-communications systems with pervasive and personal appliances, there will be a huge number of networked intelligent devices and information appliances functioning as self-organizing and managed networks (Figure 3).

In the near future, information communications systems with invisible networked devices, sensors, and appliances will transform businesses, public administration, public services, and the way we communicate within digital networks. Digital networks through the new form of information transport by ultra wideband and WiMax technologies will boost the intelligent networks development around the entire globe. The computer is becoming invisible and everywhere simultaneously. This is the beginning of the invisible computer era.

Parallel to this development is that of networking, which conceivably could result in all the invisible computers in the world being networked into a single virtual computer. This would lead to the evolution of a computer that would be everywhere and nowhere at the same time. Technology itself and on its own is not a cultural determinant. Technology is the invention of a particular culture—a cultural expression. The relationship between culture and technology is not linear and monodirectional but rather multidimensional and hyperspatial. Ten years from now, the computer as we know it today will be an anachronism, a device consigned to museums. Instead, the digital information and services once delivered via conventional computers will be available through almost everything we touch. At the heart of this next generation of computing is the network. It will be pervasive and personal. Looking out a decade or two, every person and thing could be instrumented with sensors that feed data into the content base and take actions on behalf of the client.

The new term, speckled computing, goes in that way. It offers a radically new concept in information technology that has the potential

---

**Figure 3. Scenario of universal personal information appliance**

[Diagram of a universal personal information appliance with labels for calls, clock, calculator, calendar, games, location, player camera, remote controller, always on—always online, computer, browser, email, map wallet, ID, key, multi-modal user interface, control of life(style).]
to revolutionize the way we communicate and exchange information (Arvind, 2005). Computing with Specknets will enable linkages between the material and digital worlds, and it is the beginning of truly ubiquitous computing. As the once-separate worlds of computing and wireless communications collide, a new class of information appliances will emerge. Where once they are used regularly, the post-modern equivalent might not be explicit after all. Rather, data sensing and information processing capabilities will fragment and disappear into everyday objects and the living environment. At present, there are sharp dislocations in information processing capability—the computer on a desk, the PDA/laptop, the mobile phone, smart cards, and smart appliances. However, Speckled Computing, the sensing and processing of information, will be highly diffused. The person, the artifacts, and the surrounding space become computational resources and interfaces to those resources. Surfaces, walls, floors, ceilings, articles, and clothes will be invested with a computational activity for rich interactions with the computational resources.

The information appliance is the natural outcome in the evolution of information processing. That is why they were foreseen a long time ago. Digital computers started out as expensive mainframes accessible to a few. The next step was the personal computer that individuals could own, and it was incorporated into everyday human activities. Thus, it was essential to have as much functionality in the personal computer as possible. Information technology is making feasible small and inexpensive devices that are smart. This helps to push intelligence closer to the people, the ultimate customers of information technology that accept some of the hidden cultural attributes of the entire community.

In a smart environment, the interaction endpoints simply could not be cognitively or physically visible. In essence, the user may have no idea that they are engaging in a computer-mediated communication. A smart environment is a composite space made from many individual objects. These objects are either fixed or mobile. The term invisible means that a technology has become so natural (common) and so comfortable that we use it all the time without ever thinking of it as a technology or a number of linked technologies. These invisible technologies are taken for granted. Since they are no longer technological, we can afford to think of them as customary, as the day-to-day workings of our world. As it continues to become more acceptable and as people come into the any-information system more proficient with these new tools, computer technologies will become increasingly invisible.

The exact example of the ubiquitous e-technologies environment is under way through the ubiquitous city in South Korea. New Songdo City (U-city), being built on a manmade island, will feature pervasive computer technology throughout, driven by RFID tags and CDMA wireless communication. Although many Western observers would find the lack of privacy disquieting, Asian countries are more interested in the technological potential of such environments. U-life will become its own brand, its own lifestyle. Residents will enjoy full videoconferencing calls between neighbors, video on demand, and wireless access to their digital content and property from anywhere in Songdo. At the same time, privacy is also encountered: all information services will be anonymous, and they will not be linked to user identity.

INVISIBLE E-TECHNOLOGIES’ INTERACTION WITH INVISIBLE E-CULTURE

Our relationship with the manmade world is dominated by the paradigm of the device. This paradigm creates an illusory separation between the technological means (the machinery, the medium) and the technological ends (the commodity, the message). Technological progress generally is
devoted to increasing the availability of technological goods, to make them everywhere available instantly without risk or hassle. At the same time, we want the machinery to become invisible. For various reasons, we repress ourselves from seeing the machinery and our dependence upon it. Our conscious awareness focuses on the message, and we refuse to acknowledge the medium (Pribram, 1971). Our perception of the world and our place in it are inextricably mediated through technology and the device paradigm. Our discovery of ourselves (identity through self-determination) is technical and complicated.

We are entering a new era of computing, often referred to as ubiquitous or pervasive computing. Ubiquitous computing consists of information appliances, specialized and easy-to-use devices that avoid the complexity of the computer. The future belongs to information appliances. When technology changes rapidly, greater ease of use serves to attract more users and developers, creating new frustrations. The most we can do is ameliorate the spread of the information appliance products and services. To do this, it appears necessary to recognize that flexibility and ease of use are in an unavoidable conflict and that the optimal balance between those two factors differs among users. Therefore, systems should be designed to have degrees of flexibility that can be customized for different people. Information appliances will be popular, since they will provide many new services for which the desktop computer is ill-suited and will do so in user-friendly ways. However, they will introduce their own complexity, and the level of frustration with information technology will not decrease. This is a result of the conflict between usability and flexibility. The information appliance market will be anything but mature for a long time to come. The emphasis in information processing has been and is likely to continue to be on development of novel applications. When the available information is stored on computers, it is important for information management applications to be able to model users’ interpretations of their data and to capture the possibly different meanings, semantics links, and relationships with which users associate the information units available. This is in correlation to one’s personal culture. For this purpose, various Personal Information Management tools are being developed to assist the user with navigation/browsing over various forms of personal digital data. As an example of capturing, organization, and archiving new media content, the MyLifeBits project is very explanatory. MyLifeBits has the aim to store in digital form everything related to the activities of an individual, providing full-text search, text and media annotations, and hyperlinks to personal data (Gemmell, Aris, & Lueder, 2005).

Information technology should mature to the humane technology of appliances in which the technology of the computer disappears behind the scenes into task-specific devices that maintain all the power without the difficulties. This could be explained by the technology of radio. Thus, computers should evolve the way radio receivers did (Norman, 1998). However, there is a problem of motivation, beliefs, knowledge, and frustration dealing with the information technology. This is considering culture (dominantly invisible culture). The problem is that with information appliances and by invisible technologies, we are confronted with the services that must be well-understood and stable. We will not see this scenario with information appliances, not for a long time. In a world with a huge potential in services, content, and navigation, we cannot know how people will want to use information appliances.

Careful design that is focused on human factors and incorporates powerful processors and software can provide information appliances that are a delight to use. However, once the number of devices to be connected increases and wireless communication with WiFi networks expands, the difficulties will increase. Building complicated systems that work is hard. Building ones that work and are user-friendly is much harder.
Further, it is necessary to balance the demand for user friendliness with the demand for more features. A tradeoff between flexibility and ease of use is unavoidable. The problem is that we should not be thinking just of individual information appliances. We have to be concerned with the whole system, which is likely to be complex. The problem is also how to balance flexibility and ease of use in a way that can be customized for people with different needs. This problem is especially focused on multimedia home systems. The home information appliance environment is likely to be more complicated than the office environment today. In addition, many users will be less knowledgeable about information technology than the typical office worker will (Ronfeldt, 1992). Therefore, it will be essential to outsource the setup and maintenance of home computing and electronics to experts. This notion opens new ways for information technology professionals that are accessing information appliances environment (this environment asks for new interface design, navigation methods, and computational power over networked appliances). Hence, there is convergence of culture and technology in use, be they visible or invisible.

Positive E-Technologies

Rheingold (2000) argues that the technology that makes virtual communities possible has the potential to empower ordinary citizens at a relatively small cost. E-technology potentially can provide citizens and professionals advantage and power, which is intellectual, social, commercial, and political. At the same time, civil and informed people must understand the advantages that e-technology provides. They must learn to use it wisely and constructively, as it cannot fulfill its positive potential by itself. Thus, the positive e-technologies should be developed as reduction technologies that make target behaviors easier by reducing a complex activity to a few simple steps. One of the most explicit theories that attempts to describe our natural inclination to do a cost/benefit assessment is expectancy theory. This theory posits that behavior results from expectations about what alternatives will maximize pleasure and minimize pain. E-technologies also should be self-monitoring technologies to perform tedious calculations or measurements, helping people achieve goals or outcomes. Ideally, these technologies work in real time, giving users immediate feedback on a performance or status. When people can take immediate action on a persuasive message, psychologists have found the message more persuasive than when presented at other times. The recent effort on real-time speech translation is an obvious example of these notions. Researchers from the International Center for Advanced Communication Technologies (interACT), a joint venture of Carnegie Mellon and the University of Karlsruhe, have developed a wearable system that allows real-time speech translation. The system consists of sensors that detect mouth muscle movements, translates that to a spoken language, and then retranslates that into other languages. It will make communication and cultural learning more likely, since people using this technology will be empowered to come together when they otherwise would not interact.

Culture is omnipresent in all technological advancements over the course of history, whether it is the result of intrinsic societal dynamics or the extrinsic factors of the environment. As history clearly documents, whenever technology changes, some pressing force of culture has had an effect on it. Moreover, there is a sort of invisible complimentary system between culture and technology; that is, whenever technology changes, the culture will adapt its way of life to fit the technology. For example, with the invention of the technology necessary for agriculture, cultures worldwide changed their hunting and gathering way of life in order to use the new technology and expand its horizons. This would be expected with the information technology, too. In essence, culture indeed influences human technology, but
technology also simultaneously molds the way in which cultures function.

**PERSUASIVE AGENDA**

Like human persuaders, persuasive interactive technologies can bring about positive changes in many domains, including health, business, safety, and education (Dillard & Pfau, 2002). With such ends in mind, the new area of information technology development is created under the term *captology*. Captology focuses on the design, research, and analysis of interactive computing products created for the purpose of changing people’s attitudes and behaviors. The fact that people respond socially to computer products has significant implications for persuasion. It opens the door for computers to apply a host of persuasion dynamics that are described collectively as social influence. These dynamics include normative influence and social comparison as well as less familiar dynamics such as group polarization and social facilitation.

Just as the term *software* shifts the emphasis from media/text to the user, the term *information behavior* also can help us to think about the dimensions of cultural communication, which previously went unnoticed. These dimensions always have been there, but in an information society, they have rapidly become prominent in our lives and, thus, intellectually visible. Today, our daily life consists of information activities in the most literal way: checking e-mail and responding to e-mail, checking phone messages, organizing computer files, using search engines, and so forth. In the simplest way, the particular way people organize their computer files, use search engines, or interact on the phone can be thought of as information behavior. Of course, according to a cognitive science paradigm, human perception and cognition, in general, can be thought of as information processing. While every act of visual perception or memory recall can be understood in information processing terms, today there is much more to see, filter, recall, sort through, prioritize, and plan. In other words, in our society, daily life and work largely revolve around new types of behavior activities that involve seeking, extracting, processing, and communicating large amounts of information. Information behaviors of an individual form an essential part of individual identity. They are particular tactics adopted by an individual or a group to survive in information society. Just as our nervous system has evolved to filter information existing in the environment in a particular way that is suitable for information capacity of a human brain, so we evolve particular information behaviors in order to survive and prosper in an information society. In today’s world of information, people suddenly are shifting their attention to the Web for their computing needs.

Levy (1998) contends that communication in the virtual world can cultivate collective intelligence, which can encourage the development of intelligent communities. He states that sharing of information, knowledge, and expertise in e-communities can promote a kind of dynamic, collective intelligence, which can affect all spheres of our lives. He contends that the virtual world can foster positive connections, cooperation, bonds, and civil interactions. In e-groups or e-communities, which are flexible, democratic, reciprocal, respectful, and civil, this collective intelligence can be enhanced continually. Researchers in science, education, business, and industry are pooling their collective intelligence, knowledge, and data in collaborative memories. These are virtual centers in which people in different locations work together in real time, as if they were all in the same place. Science, education, commerce, and industry have become increasingly global. Collaboration, which is efficient, maximizing, and timesaving among distance researchers in these fields, has become more critical.

This new e-culture paradigm within the Web users’ community opens the ways for Web 2.0
Information-Communications Systems Convergence Paradigm

e-technologies platform comprising the set of principles and practices based dominantly on the user behavior and cultural values of collaboration. Most users find that Web 2.0 sites are extremely useful, because they are always available (whenever they need it and anywhere they go) with their information. Web 2.0 is the network as platform, spanning all connected devices. Web 2.0 applications are those that make the most of the intrinsic advantages of that platform:

- Delivering software as a continually updated service that gets better, the more people use it.
- Consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others.
- Creating network effects through architecture of participation.
- Going beyond the page metaphor of Web 1.0 to deliver rich user experiences.

One of the key aspects of Web 2.0 is that it connects people so that they can participate effortlessly in fluid conversations and dynamic information sharing. At the same time, information appliances and computing devices are giving people permapresence on the Web. Before now, the user consciously had to go to cyberspace by sitting at a desktop computer and looking at it through a display. Web 2.0 applications will become invisible as they become more popular, and there also would not be such a phrase as “going on the Web.” Moreover, if the network is omnipresent and invisible, we do not need the term cyberspace anymore. Web 2.0 is also more human and a social one labeled with social interactions like conversation, sharing, collaboration, publishing, which could be supported by the corresponding processes (blogging, tagging, sharing, publishing, networking) and content formats (blogs, wikis, podcasts, folksonomies, social software). In addition, Johnson and Kaye (2004) stated that the Web would become a trustworthy place and the users would take it with much more reliance and credibility.

Computers Influence our Thoughts and Actions

Although culture is mostly learned, it is bound by necessity to a particular setting or context of its behavioral and material articulation. Culture is both conservative and adaptable. Culture is articulated symbolically and has the function of symbolically integrating the diverse moments and spaces of culture into a coherent sense of order. This format is emerging throughout the social field as a format of technology (the point-to-point Internet, file sharing, grid computing, blogs), and as a third mode of production producing hardware, software (often called open sources software) and intellectual and cultural resources (wetware) that are of great value to humanity (GNU/Linux, Wikipedia).

Cognitive scientist Clark (2003) believes that we are liberating our minds, thanks to our penchant for inventing tools that extend our abilities to think and communicate, starting with the basics of pen and paper and moving on to ever more sophisticated forms of computers and e-technologies. He declares that we are, in fact, human-technology symbionts, or natural-born cyborgs, always seeking ways to enhance our biological mental capacities through technology. The persuasive e-technologies are in front of us to solve the problem of difficulties in utilization of the computer, which complexity is fundamental to its nature. We have to start over again to develop information appliances that fit people’s needs and lives (Norman, 1998). In order to do this, companies must change the way they develop information system products. They need to start with an understanding of people: user needs first, technology last. Companies need a human-centered development process, even if it means reorganizing the entire company.
People are more readily persuaded by computing technology products that are similar to them in some way (Fogg, 2003). Although people respond socially to computer products that convey social cues, in order to be effective in persuasion, hardware and software designers must understand the appropriate use of those cues. If they succeed, they make a more powerful positive impact. If they fail, they make users irritated or angry. With that in mind, when is it appropriate to make the social quality of the product more explicit? In general, it is appropriate to enhance social cues in leisure, entertainment, and educational products, especially with smart mobile devices (Rheingold, 2002). Users of such applications are more likely to indulge, accept, and perhaps even embrace an explicit cyber social actor. When perceived as social actors, computer products can leverage these principles of social influence to motivate and persuade.

**DISCUSSION ON NEW FORMS OF CULTURAL RESPONSIBILITY OF IT PROFESSIONALS**

Culture and ethics are a very important part of our everyday life in information society. The invention of new e-technologies tends to bring many different dilemmas into the lives of the creators and the people who use them. Some technologies have been created without choice, and we must make sure we fully understand how to use them properly. The introduction and use of new technologies require a check against the moral structure of the society and the ethical beliefs of the individuals that will feel the effects of such an addition to their lives (Postman, 1992). This belief should be the foundation of innovation so that members of the society can have a strong, viable, and ethical solution to satisfy their wants and needs and to extend their capabilities. Ubiquitous computing and smart environments will be characterized by massive numbers of almost invisible miniature sensing devices that potentially can observe and store information about our most personal and intimate experiences. The new forms of direct access to information and knowledge create new forms of e-culture. It concerns not only users but the community of information professionals as well.

Technology can be a powerful tool for change, especially when used responsibly. Responsible IT management should be an important part of any socially responsible enterprise’s strategies, policies, and practices. Users should have information technology choices that can and should reflect organizational, community, and national values and social responsibility. These notions are considering IT professionals’ activities; they should create applications that guarantee accessibility. Accessibility to information via the information-communications systems should not be inhibited by disability or resource limitations, and the design solutions should be for the user experience. Usability of information technology solutions requires attention to the needs of the user (information consumer). The information-communications systems create new psychological demands from human. They ask us to bring a greater capacity for innovation, self-management, and personal responsibility. They also demand social responsibility of information technology professionals. Information technology is the wave of today and the future. Society must adapt to it by creating responsible rules, norms, ethics, and knowledge workers that will enhance its rapid growth.

Many firms acting on the global scene via information-communications systems are committed to incorporating socially responsible projects into their policies and activities. Corporate social responsibility is a development that is here to stay for the long term as a part of corporate policy influencing the company’s involvement in the well being or development of local as well as global communities (Furnham & Gunter, 1993). Information technology firms are in a unique
position to distribute their high-tech expertise and cultural values in development projects. The new information society environment poses a new relation between values promoted by the new digital civilization and the traditional moral and cultural values created by the major world civilizations. Computer technology and ethical egoism are the products of secular research within a free market capitalist society. The majority of non-Western societies and some Western, as well, follow ethical rules created within traditional culture. These rules are centered on guiding the individual in properly fulfilling his or her role within the society, which means the superiority of the society over the individual. The changes that information technology is bringing to people’s lives are revolutionary, and one of the features of every revolution is that it is at the same time both a process of creation and of destruction. The revolutionary process itself is a very rapid one, which means that there is little or no time for a methodical and deep reflection on it while the process is actually in progress. These points ask for more attention from information technology professionals to cope with the culture exposed through the visible objects. They also should implement invisible culture elements when designing new information services. One possible way of minimizing the harm could be through incorporating the experiences of the process of intercultural dialogue into the process of creating a global e-culture of the information society.

REFERENCES


*This work was previously published in Information Resources Management: Global Challenges*, edited by W. Law, pp. 54-74, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 7.10
IT Artefacts as Socio–Pragmatic Instruments: Reconciling the Pragmatic, Semiotic, and Technical

Göran Goldkuhl
Linköping University & Jönköping International Business School, Sweden

Pär J. Ågerfalk
University of Limerick, Ireland, & Örebro University, Sweden

ABSTRACT

There are many attempts to explain success and failure in information systems. Many of these refer to a purported sociotechnical gap. In this article we develop an alternative approach that does not impose such a strong dichotomy, but regards social and technical rather as dimensions along which to study workpractices. The developed theory involves not only the “social” and “technical” constructs, but also other generic ones, namely “instrumental,” “semiotic,” and “pragmatic.” We call this theory socio-instrumental pragmatism. To illustrate the theoretical concepts introduced, we use an example brought from an extensive action research study including the development of an information system in eldercare, developed through a participatory design approach.

INTRODUCTION

Development and implementation of an information system (IS) is a very demanding task, and many times the expectations from such endeavours are not met. Unexpected negative effects often arise, while anticipated positive effects fail to appear. There are many attempts to explain IS failure (and, indeed, success) in general terms. Some of them refer to a sociotechnical gap — a gap between what is socially required and what is technically feasible (e.g., Ackerman, 2000). Such explanations tend to make a sharp differentiation between the social and the technical. For example in the sociotechnical tradition represented by Mumford and Weir (1979), there are discussions about balancing the technical system and the social system. This is built upon a view that
computerised information systems are technical systems with social and organisational effects—a view that seems almost entirely to permeate mainstream IS research (e.g., DeLone & McLean, 1992, 2003; Benbasat & Zmud, 2003). This is also in line with the soft systems view that there is a “serving system” to support a “system to be served” (Champion & Stowell, 2002). There are criticisms toward such a conceptualisation. For example Nurminen (1988, p. 82) writes: “By removing the social dimension from the systems entity, we imply that the technical system is basically non-social.” In the same spirit, Goldkuhl and Lyytinen (1982) suggest that the traditional view of information systems as “technical systems with social implications” should be inverted to “social systems, only technically implemented.” As pointed out by Mead (1934):

“Language does not simply symbolise a situation or object which is already there in advance— it makes possible the existence or appearance of that situation or object, for it is part of the mechanism whereby that situation or object is created.”

Since every IS uses language for purposes of communication and understanding (Goldkuhl & Lyytinen, 1982), what Mead claims about language also counts for information systems.

Instead of a separation into a social realm (humans acting in the IS environment) and technical realm (the IS), another approach is proposed here: using “social” and “technical” as dimensions along which to study workpractices. The theoretical way to proceed is to articulate a common theory for both the IS and its organisational context. The concepts of social and technical are however not found to be sufficient. The purpose of this article is to outline a theory appropriate for interpretation, description, explanation, and evaluation of the interaction between information systems and their organisational context. The developed theory involves not only the “social” and “technical” constructs, but also other generic ones, namely “instrumental,” “semiotic,” and “pragmatic.” As we shall see below, these constructs are “generic” in the sense that they are not specific to any particular empirical setting, but are high-level categories useful in describing and discussing social action in relation to information systems in general. We call this theory socio-instrumental pragmatism, aligning with the work of Goldkuhl and Ágerfalk (2002) and Goldkuhl and Röstlinger (2003).

The article proceeds as follows. In the next section we briefly introduce some concepts of socio-instrumental pragmatism. To illustrate the use of socio-instrumental pragmatism as a theory of information systems, we use a simple example of an IS in the subsequent section. The example is brought from an extensive empirical study. This study involves an action research endeavour, including development of an IS in an eldercare setting. The IS and its supported workpractice were developed through a participatory design approach. We do not describe this case study in any detail, but use part of the developed system and the workpractice to illustrate our theoretical endeavour.

We then condense our conceptualisation of the IS and its organisational context in the following section, where important concepts are clarified and related to each other. Our contribution should be understood as a way to conceptualise the information technology (IT) artefact and its context. Hence, this article can partly be seen as a response to the requests for theorising the IT artefact as espoused by Orlikowski and Iacono (2001) and Benbasat and Zmud (2003). Those papers have given rise to quite a debate (e.g., Alter, 2003; Galliers, 2003). It is beyond the scope of our article to directly engage in this debate, although near the end, we comment on our contribution as a response to these requests for theorising the IT artefact. The article concludes with a brief summary of the main points.
SOCIO-PRAGMATIC PRELIMINARIES

Every attempt to theorise on information systems and their organisational uses relies on some ontological assumptions. Are there any fundamental entities and processes in the social and artificial world that should direct our conceptualisations? The basic stance taken here is pragmatic. Blumer (1969, p. 71) states “that the essence of society lies in an ongoing process of action — not in posited structure of relations. Without action, any structure of relations between people is meaningless. To be understood, a society must be seen and grasped in terms of the action that comprises it.”

With inspiration from Blumer and his pragmatic forerunners Dewey (1931) and Mead (1934), we will let actions appear as the ontological core in our analysis. In the IS field, such a position has been argued by Goles and Hirschheim (2000), guided by discussions on the “new pragmatism” (Wicks & Freeman, 1998).

The kinds of actions we are interested in are social actions — that is, actions performed in a social context. Weber (1978, p. 4) made a classical definition of social action: “That action will be called ‘social’, which in its meaning as intended by the actor or actors, takes account of the behaviour of others and is thereby oriented in its course.”

Our interpretation of this definition is that a social action (performed by an actor) has social grounds (“takes account of the behaviour of others”) and social purposes (“thereby oriented in its course”).

In communication and other direct interaction between actors, the social character is obvious. A communicative act (e.g., a question) functions as an impulse or initiative for a subsequent act (presumably an answer), which then is seen as a response to the first act (Sacks, 1992; Linell, 1998). Talk and other social interaction are seen as such chains of initiatives and responses.

The social grounds not only comprise the direct intended influence toward the actor; when acting, the actor may use immaterial and/or material instruments. These instruments usually have a social origin, and hence the instrumental use in action implies a social influence on the actor (Vygotsky, 1962; Wertsh, 1998). This is obvious when people use language. All use of language is social in the sense that it uses the social instrument of language. The actor’s thinking, speaking, and writing are shaped by the socially constructed categories of the world manifested in language or other sign systems (Berger & Luckmann, 1967). This also means that the use of an IS must be social, since there can be no such use without the exploitation of signs.

People intervene into the world through their actions. Actions are intended to make a difference (Dewey, 1931). When one usually talks about actions, such intervenient actions are implied. However, all human action is not intervenient and overt. There are covert actions, which are not intended as external changes. The perception and interpretation of the world is also seen as actions, although as covert ones (Schutz, 1970; Strauss, 1993). This also involves thought and conscious deliberation through reflection. Such covert actions (interpretation and reflection) do not aim at external change, rather a change to the actors themselves in their knowledge about the world. Social interaction consists of related intervenient and receptive actions. An ordinary communicative situation with a conversation between two actors consists of continual acts of speaking (i.e., intervening) and listening (i.e., receiving) (e.g., Clark, 1996). The production of an utterance is followed by an interpretation of that utterance by the other actor. Conversation consists of such reciprocal pairs of intervention and interpretation (Goldkuhl, 2003), which are embedded in “adjacency pairs” of communicative actions (Sacks, 1992), for example a question and a corresponding answer (mentioned above).

Such acts in communication processes are semiotic acts, since they specifically are dealing
with signs (i.e., producing signs and interpreting signs). People perform other kinds of acts as well. They act toward material objects with the intent to accomplish material changes in the world. Most material actions are also social actions, since they have social grounds and social purposes as stated above (also see Goldkuhl, 2003). Material actions are often parts of social interactions, and as such they are intertwined with semiotic acts (Goldkuhl, 2003; Andersen, 1990). Vološinov (1985) expresses this in a clear way:

“Verbal communication can never be understood and explained outside of this connection with a concrete situation...In its concrete connection to a situation, verbal communication is always accompanied by social acts of a nonverbal character, and is often only an accessory to these acts, merely carrying out an auxiliary role.”

In material actions, people often use external instruments (e.g., IT artefacts) to enable or facilitate their actions. Also, in communication, instruments are often used to improve the transfer of messages. Instruments such as telephones, fax machines, and computers make communication possible over time and distance. Instruments play different roles in action and interaction; they can also be more or less advanced. Goldkuhl and Ågerfalk (2002) and Goldkuhl and Röstlinger (2003) have distinguished three levels of instruments:

1. Static tool (artefact-supported human action)
2. Dynamic tool (human-artefact cooperative action)
3. Automaton (human-defined artefact action)

A human must use a static tool directly in order to make a difference. For example, an axe is used to chop wood. A human uses a dynamic tool in a cooperative way. For example, a car must be manoeuvred when driven. Automatic machines can perform work independently according to its programme, after being initiated by a human. When started, a washing machine works by itself.

The most central of these socio-pragmatic preliminaries are summarised and visualised in the generic model of social action depicted in Figure 1. More elaborated descriptions can be found in Goldkuhl and Ågerfalk (2002) and Goldkuhl and Röstlinger (2003).

When acting through and by means of an IS, the IS is used as the instrument by which intervening actions, and possibly also receptive...
actions, are performed. In the next section, this socio-pragmatic view of sociotechnical action is elaborated further by means of examples from a case study.

**ACTIONS AND INSTRUMENTS IN WORKPRACTICES: AN ELDERCARE CASE STUDY**

For the purpose of illustration, we will use material from a case study — a longitudinal action research-based case study — on the development of a municipal eldercare unit. The action research project consisted of integrated workpractice and information systems development. A new IS was developed in a participatory design setting, where researchers and eldercare personnel participated in a co-design endeavour. Data generation in the case study has been pursued with different qualitative research methods: interviewing directors and nursing assistants, observing their work, collecting and analysing several documents, and participating in development seminars. A closeeness to the empirical phenomena was necessary in order to gain reliable data. Experiences from the case study have been reported previously by Cronholm and Goldkuhl (2002) and Goldkuhl and Röstlinger (2002).

Let us start the discussion with the eldercare practice before computerisation. The eldercare practice consists of nursing assistants giving care to elderly people living in their own residences. The elderly people need assistance with ordinary tasks like hygiene, dressing, and cleaning, as well as simple medical attention on a daily basis. A nursing assistant visits each client on one or more occasions each day. The visits are regulated by daily schedules, which inform the nursing assistant what tasks to perform. There exist different schedules depending on the kinds of tasks to perform and the time of day. There are typed contents on the schedules, but they also consist of handwritten annotations. Besides the schedules, there are also more informal communication between the nursing assistants, such as handwritten notes and verbal interaction. Quality assurance problems were encountered in the eldercare practice which gave rise to the development of an IT system to support communication and documentation. Schedules are nowadays mediated by the IT system, and this has improved the quality assurance and the individualisation of the eldercare which are important objectives of the workpractice.

In order to understand what the new IT artefact does, we need to understand its role in the workpractice. In order to do this, let us focus on three different types of actions in the workpractice:

1. production/reproduction of daily schedules,
2. reading of the schedules before home care visits, and
3. care service provided to the elderly clients.

Based on the socio-pragmatic perspective outlined above, we focus on actions as a kind of basic unit of analysis. A workpractice consists of actions, actors, signs, and material artefacts related in meaningful ways (Goldkuhl & Röstlinger, 2002, 2003). The way social practices appear as meaningful to an inquirer is through making their constituent actions become visible — what people do in the workpractice.

The first two kinds of action (creating and interpreting the schedule) are two inter-related parts of a communication process. One nursing assistant may, for example, annotate something to the schedule at one occasion as a message to her colleagues (“remember to take out the laundry”). Another nursing assistant reads the note when she is about to visit an elderly resident. And she takes this into account during her visit to the resident (taking care of the laundry). The first two actions are actions dealing with language (writing and reading) and are thus parts in a communication
process. The last action is not a linguistic action. It is a material action—changing physical objects in the world. This reasoning may be illustrated by the model of social action depicted in Figure 1.

We distinguished above between intervenient and receptive actions. Actions 1 and 3 in the examples are intervenient actions. Action 2 in the example is an interpretive action. All three actions are social actions, although they might all be performed in solitude without any direct interaction with another human. All actions described have social grounds and social effects. The first nursing assistant might write her note without any other colleague around her, but the communicative action is, of course, directed toward and intended for another human—that is, asking the (other) nursing assistant to visit the elder. The second nursing assistant may perform the second act—the reading of the schedule and the accessory note—in solitude. As an interpretive act it is a typical social act. It certainly has social grounds through the written communication directed toward her. This act also has social purposes. The nursing assistant reads the message in order to arrive at an informed readiness on the eve of her visit to the elder. The third act—the material act of taking out the laundry at the elder’s residence—may also be performed in solitude. This material act is however also a social act, since there are obvious social grounds and social purposes. There was a request directed to her from her colleague (i.e., social grounds). She is taking out the laundry in order to help the elder and fulfill the obligations of the eldercare centre (social purposes).

All these genuinely social actions are performed by use of different tools, or instruments. The reading of the schedule requires the schedule being presented in a readable form, and annotating it requires a writable form. Both the “pen and paper” used initially and the subsequently introduced computerised IS afford these actions. In both cases, the instrument is used for interpreting and expressing semiotic results as a basis for and as a result of social action. The first nursing assistant is guided by communicative intent. She wants to say something to her colleague on the next work shift. The computerised schedule is an instrument for her to perform this communicative action. The second nursing assistant wants to be informed before she embarks on her visits to the elderly people. She uses the computerised system comprising schedules in order to become informed. Thereby she becomes capable of performing subsequent knowledgeable actions for the clients. The nursing assistant manoeuvres the IT system in order to arrive at a specific schedule (a specific client at a specific date and work shift). This kind of manoeuvring of the IT artefact implies both technical and semiotic knowledge. The nursing assistant must of course also have relevant organisational knowledge. She must be aware of the obligations of the eldercare centre and herself in her organisational role. She must know the institutionalised way of communicating expected tasks (i.e., the work schedule), which now is included in the new IT system.

One of the main driving forces of this computerisation was to ensure the quality of the communication and documentation around the care of the elderly. The directors of the eldercare centre wanted the new IT system to be the main medium for communication. Earlier there was a mix of several different handmade forms, small handwritten messages, and oral communication. The idea was now to channel most of the communication through the IT system. The work schedule should describe all planned (desired) measures to be taken for a particular elder at one visit. This includes both standard measures generally performed for this elder and specific measures needed at one or a limited amount of occasions. This means that the work schedule in the IT system is a collection of several communicative acts performed by several persons in the eldercare staff. One role of the IT system is thus to collect and integrate these different communicative acts in a proper way—that is, to enable communications
IT Artefacts as Socio-Pragmatic Instruments

from many persons to the one nursing assistant to perform a particular visit.

Even if the IT system is considered to be the main medium for communication of work tasks, there will of course remain informal channels beside the system. There will be situations when it is more appropriate to still communicate in an informal way. The system will not hold a complete description of the elders and their different needs. Unanticipated discourses concerning the elders will be required, which are not suited for channelling through the IT system. Also, the system cannot communicate all necessary measures to be taken. The nursing assistants must of course be attentive during their visits and, based on observations and demands from the elderly, perform appropriate unplanned measures.

As shown above, the role of the IT system is to facilitate communication. The system must support and enable the eldercare staff to perform important communicative actions through the system, like work schedule planning. There are other communicative tasks not yet mentioned which are important in the workpractice. Important incidents regarding the elderly must be reported into the care journal of each elder. The IT system supports the reporting of incidents into the care journal. The system must automatically manage the care journal and on request present parts of it to those privileged to read the journal. The organisational role of this care journal is to be part of a workpractice memory. Actors in the workpractice must be able to remember what has happened and what actions have been performed. The IT system gives institutionalised support for recollecting things that occurred and were performed.

In general, workpractices are equipped with a diverse range of instruments for actors to use in their social endeavour. These are often technical instruments, such as computers and washing machines. The usefulness of these instruments within a practice is contingent upon the meaning attached to them by the actors. From this perspective, computers are mainly means to improve communication (Flores, 1998). Indeed, IT artefacts are technical instruments. Their main functioning within a workpractice, such as eldercare, however, is as instruments for social action.

This means that we can view IT artefacts in a workpractice as: 1) technological artefacts with physical properties, 2) semiotic artefacts affording communication and interpretation, and 3) social instruments used to responsively express actors’ beliefs, values, and intentions. These three aspects of IT artefacts may be analytically distinguished. It is important to see, though, that IT artefacts are not simply isolated technical systems related to a social practice, in whatever intricate way. Rather, their physical properties are what enable and restrict possible semiotic interpretations and expressions required to form the social practice at hand. To paraphrase the quotation from Mead (1934) above: they are not simply objects which are there — they make possible the existence or appearance of that situation, for they are part of the mechanism whereby that situation is created.

With this theoretical and empirical backdrop, the following two sections explore how information systems (as IT artefacts) can be conceptualised following socio-instrumental pragmatism, and how such a conceptualisation contributes to our understanding of IS theory in general.

CONCEPTUALISING SOCIO-INSTRUMENTAL ACTIONS AND SOCIO-PRAGMATIC ARTEFACTS

Information systems (as IT artefacts) are technical systems. This is obvious. They consist of hardware and software. They are technical instruments, usually not aimed for direct support of material action. Instead they support communication and other types of information handling. As such
they are communicative and semiotic devices. Information systems consist of representations such as texts and other signs.

This means that an IS is not just a restricted technical tool. An information system is a socio-pragmatic instrument. As such an instrument, it is utilised to perform social actions. These social actions are communicative in nature. An IS must thus be a semiotic instrument. It must have capabilities to process signs. In doing this in a sophisticated way, information systems rely on advanced technical equipment. This technical equipment needs to be managed by a human being who wants to utilise its semiotic capabilities. An IS is an instrument for producing messages to other people and an instrument to be informed by other people. The technicalities of an IS are necessary material bases of its functionality as a socio-pragmatic instrument. Without hardware and software, one could not write and read through an IS. The material nature of an IT artefact can make it instrumental for social purposes.

Usually, an IS has a pre-defined set of communicative possibilities as defined by its functionality and vocabulary (the latter usually defined by its database schema); not everything can be said. An IS is also an instrument for getting informed by others; otherwise it would not be a communication instrument. It must support both parts of the communication process — to express and to interpret. Information systems have the capability to execute communicative actions according to its predefined action repertoire (determined by its programmmed software). The action repertoire includes affordances for communicating and interpreting. An IT system as a communicative instrument supports writing and reading; in other words, it is a mediator of communication. IT systems usually have advanced capabilities of a mediating agent. Such a system does not only mediate one message from one person to another person. It has capabilities to mediate communication from many persons to many persons in intricate and sophisticated

Figure 2. An information system as a socio-pragmatic instrument in a workpractice
ways, including the maintenance of appropriate workpractice memory (i.e., a database). An IS as part of an organisational context is depicted in Figure 2. This figure is based on the general model of social action (Figure 1). In this socio-pragmatic view of information systems, the most important property of such artefacts is referred to as their actability. In order to be socially useful, information systems must be actable for their users (Goldkuhl & Ågerfalk, 2002; Cronholm & Goldkuhl, 2002; Ågerfalk, 2004).

Such capabilities make the IT system an organisational agent. An agent is someone who does something on behalf of someone else. Humans working in an organisation are agents for that organisation. They act as representatives for that organisation, on behalf of the organisation (Ahrne, 1994; Taylor & Van Every, 2000; Goldkuhl & Röstlinger, 2003). An IT artefact (being an organisational agent) does possess some action capabilities, but lacks typical human attributes (consciousness, intention, emotion, social awareness, empathy). In some situations it is appropriate to view the IT artefact as a mere instrument (a static or dynamic tool) to be used by humans. In other situations it may be appropriate to bring its agent capabilities to the fore — and hence its possibilities to interact with humans as other organisational agents.

One key feature of this theory (socio-instrumental pragmatism) is that it should be used both for understanding the IT artefact and its human and organisational workpractice, which it is part of. This is a kind of seamless theorising, where we do not shift theoretical perspective when moving between the artefact and its human and organisational context (Goldkuhl & Röstlinger, 2003). According to this theory, actions are multifunctional; performing an action does several things. An action can at the same time be:

- a response to an action made by someone else (i.e., socially responsive);
- an expression of subjectivity, a utilisation of immaterial instruments (such as knowledge and language);
- a utilisation of external instruments (e.g., technical artefacts);
- a compliance to organisational norms (e.g., role expectations);
- a production of action results (semiotic or material objects);
- a realisation of values and intentions; and
- an attempt to influence someone else (i.e., exerting power).

Being multifunctional, actions are also multi-dimensional. Actions performed by a human user, when acting through an IT artefact, can be described as social (social grounds and social purposes), semiotic (using and producing signs with communicative intent), and technical (managing some technical device) at the same time. When we talk about an IS as a socio-instrumental pragmatic system, we emphasise its pragmatic features and purposes. Instrumental (in “socio-instrumental”) should be read as both semiotic (language as instrument) and technical (material artefact as instrument). It is thus not sufficient to talk about an IS as a sociotechnical instrument. If we do that in a restricted way, then we may forget its primary role of being a semiotic device for organisational communication. Socio-instrumental pragmatism provides IS designers with a pragmatic view to direct attention to the whole of the workpractice and to its constituent parts of human, organisational, communicative, and technical characters.

A WAY TO THEORISE THE IT ARTEFACT

Based on a review of the complete set of 1990-1999 articles published in Information Systems Research (N = 188), one of the leading IS scholarly
journals, Orlikowski and Iacono (2001) issued a general call to “theorising the IT artefact.” The reason for taking such a measure was that a stunning 25% of the reviewed articles focused on other aspects than IT, referring to IT only incidentally or as background information. This should be understood in light of an ongoing quest for the identity and legitimacy of the field of Information Systems that started in the 1980s (e.g., Banville & Landry, 1989) and recently culminated in a sometimes heated debate caused by Benbasat and Zmud’s (2003) already famous paper on “defining and communicating the discipline’s core properties.” The question underlying these efforts of organising the IS field seems to be threefold.

First, as pointed out by Benbasat and Zmud (2003), there is the issue of what to include in IS research — what are the core properties that must be considered in order to understand the role of IT in social settings? Second, the flip-side of Benbasat and Zmud’s (2003) coin is the question of what should be excluded from IS studies — that is, what properties of social life are not directly related to IT and thus better left to other disciplines to scrutinise? Finally, we need to understand the very concept of IT — to theorise the IT artefact. We must not regard IT as a technical black box, but recognise what specific properties of this technology contribute to our understanding of its role in everyday life; “We will need to stop taking IT artefacts for granted and begin to take them seriously enough to theorise about them” (Orlikowski & Iacono, 2001).

It is generally believed that IT must be understood within a surrounding social context (e.g., Benbasat & Zmud, 2003). It is also generally believed that a fruitful analytic approach is to separate IT from non-IT, technical from social, and then to study how the two entities interact. This preferred separation of concerns is obvious from Benbasat and Zmud’s (2003) model of the “IT artefact and its immediate nomological net,” which includes the three core IS properties: the IT artefact, usage, and impact. It is also obvious from DeLone and McLean’s (1992) IS Success Model, which includes the corresponding variables of system/information quality, use, and impact.

We argue that in order to seriously theorise the IT artefact, we need to develop theories that are useful in describing IT artefacts, in describing IT artefacts’ social context, and in describing IT artefacts contextually. We need theories that let us see the social in the technical and the technical in the social. Obviously, this view is not too far from that of the actor-network theory or ANT (as in Latour, 1991), where technology and people are both seen as social actants. However, as pointed out by Orlikowski and Iacono (2001), being concerned specifically with IT, we cannot accept general sweeping notions of technology. We must acknowledge the specific characters of IT, which include their symbol processing (semiotic) properties, their tool (instrumental) properties, and their social (agency) properties. The socio-instrumental theory of IT artefacts outlined in this article allows us to use one theoretical lens to study information systems in concert with humans, without losing our focus on core IS properties and without black-boxing technology, seen as something that is simply a cause or effect of something else. It enables us to view both humans and IT artefacts as organisational agents, and to do this without losing sight of their fundamental differences in agency. This prevents us from falling into the trap of a too symmetrical view of humans and IT artefacts, which is one main criticism toward applications of the actor-network theory within the field of IS (e.g., Walsham, 1997). The symmetrical view of ANT should thus not be confused with our ideal of seamless theorising mentioned above.

Orlikowski and Iacono (2001) conclude their research commentary by offering five premises for theorising IT artefacts. In short, they argue that:
IT Artefacts as Socio-Pragmatic Instruments

1. IT artefacts are always value based, since they are designed and used by people with different interests, values, and assumptions.
2. IT artefacts are always embedded in a historical and cultural context.
3. IT artefacts are usually made up of several components, and generic terms such as “the Internet” and “the Technology” should generally be avoided.
4. IT artefacts are not fixed and independent, but co-evolve with the workpractice in which they are embedded.
5. IT artefacts are subject to changes in technology.

Arguably, these premises go well with the proposed socio-instrumental pragmatism, which may thus be regarded as an instance of the sort of theorising called for by Orlikowski and Iacono (2001). On the other hand, we think that Orlikowski and Iacono (2001) might be overly discouraging when stating that: “Given the context-specificity of IT artefacts, there is no single, one-size-fits-all conceptualisation of technology that will work for all studies.” Indeed, any workpractice-related theory is by definition localised to that specific workpractice. On the other hand, we would like to argue that there are generic conceptualisations and units of analysis, such as those put forth in this article, that are general in nature but pragmatic enough to be applicable in most studies. There are a set of “core properties,” or aspects, of IT in workpractices that we should pay attention to in order to understand any IS in practice. We have termed these aspects the social, the technical, the instrumental, the semiotic, and the pragmatic. These properties, in turn, are by nature situational and can as such assist in eliciting important aspects embedded in particular research settings.

CONCLUSION

It is generally agreed that in order to understand IT and its functioning in workpractices, the surrounding social context must be seriously taken into account. A common view of conceptualising IT and its use is to make a sharp distinction between the technology as such on the one hand, and social prerequisites and effects on the other. Contrary to this popular belief we have articulated and illustrated a theoretical perspective that allows for conceiving of technological aspects and social aspects of workpractices in an integrated way. The perspective, which we have termed socio-instrumental pragmatism, is founded in pragmatic theories of social action, and the main unit of analysis is that of mediated social action. By viewing the use of information systems as the performance of technologically mediated social actions, we can distinguish between three important roles of IT: as an affordable technological base, as mediator of semiotic signs, and as a social instrument to responsively express actors’ beliefs, values, and intentions. Instead of separating a social realm from a technical realm, we use “social” and “technical” as dimensions along which to study workpractices. This way of reconciling the pragmatic, social, semiotic, and technical allows for a uniform approach to studying and understanding the relationship between IT and humans by directing attention to humans’ social actions performed through and by means of IT. This approach goes beyond sweeping general notions of technology as part of socially constructed reality, and allows us to explore and exploit the genuine character of IT as part of an organisational agency.

ACKNOWLEDGMENTS

This work has been financially supported by the Swedish Association for Local Authorities,
the Swedish Agency for Innovation Systems (VINNOVA), the Science Foundation Ireland Investigator Programme, B4-STEP (Building a Bi-Directional Bridge Between Software Theory and Practice), and the Knowledge Foundation (KK-Foundation) through the programme for the promotion of research in IT at new universities and university colleges in Sweden (IT-Lyftet). The authors would like to thank Jack Downey for useful comments on a draft version of this article.

REFERENCES


*This work was previously published in the International Journal of Technology and Human Interaction, edited by B. C. Stahl, Volume 1, No. 3, pp. 29-43, copyright 2005 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 7.11

Ignorance is Bliss: The Effect of Increased Knowledge on Privacy Concerns and Internet Shopping Site Personalization Preferences

Thomas P. Van Dyke
The University of North Carolina Greensboro, USA

ABSTRACT

Studies have shown that people claim that privacy matters to them but then they often do things while browsing that are risky in terms of privacy. The seeming inconsistency between professed privacy concerns and risky behavior on the Internet may be more a consequence of ignorance rather than irrationality. It is possible that many people simply don’t understand the technologies, risks, and regulations related to privacy and information gathering on the Web. In this study, we conducted an experiment to determine the answer to the following question: If people understood the risks and technology, would that knowledge alter their level of privacy concern and their preferences concerning e-commerce Web site personalization? Results indicate that increased awareness of information gathering technology resulted in significantly higher levels of privacy concern and significantly reduced preferences for Web site personalization. Implications of the findings are discussed.

INTRODUCTION

Individuals are willing to participate in diverse activities online—from e-mailing friends and looking up personal medical information to purchasing a wide variety of goods and services. While consumers benefit from their activities online, businesses also benefit from information gained while consumers browse. The Internet environment allows businesses to collect and analyze more personal information with greater ease and efficiency than ever before. Firms can use several methods to collect information about visitors to their sites. These include overt methods, such as registration forms, Web surveys and order forms, as well as covert methods including spyware, Web bugs and cookies. The information gathered serves...
as an important input into marketing, advertising, customer service and product-related decisions by online firms. The information gathered also allows firms to offer personalization (i.e., mass customization) to the Web site. This has the potential to benefit both the customer, through added convenience and options, as well as the firm by encouraging increased sales.

However, the consequences of information gathering are not all positive. The ability of firms to gather so much information creates the risk of possible misuse and generates concerns over information privacy among users. These privacy concerns impede e-commerce. The Federal Trade Commission estimates that online retail sales were reduced by $18 billion due to privacy concerns in 2002 (Gellman, 2002).

Users claim that privacy is important to them (Westin, 2003). However, they are constantly taking actions that are risky in terms of privacy. Internet users are often required to make tradeoffs, taking actions that sacrifice privacy in return for convenience, such as Web-site personalization. These actions often appear to be in contradiction of their professed attitudes regarding their personal privacy.

For example, Internet users have consistently indicated that they did not want firms to track their Web surfing habits (Westin, 2003). However, people routinely accept cookies through their Web browsers by default. According to one study, users rejected fewer than 1% of cookies in over a billion page views (Websidestory, 2001).

There are several possible explanations for the seeming contradictions between user attitudes and actions. According to Kristol (2001) these include:

- Users know how cookies can collect information and track them but are unconcerned.
- Users don’t know how cookies can be used to track them.
- Users have inadequate means to select which cookies to accept so they just give up and select them all.
- Users assume that the firms collecting the information will protect it and use it discreetly (not true in all cases).
- Users assume (incorrectly) that that they are protected by governmental regulations that will prevent Web sites from misusing information about them.

Determining which of these explanations are true regarding users has important policy implications. Companies and interests groups that benefit from the information collected on the Net back self-regulation, and that is the current model used in the United States. Research by authors such as Westin (2003) can be used to support self-regulation on the grounds that people are free to make a choice. According to surveys, the majority of Internet users fall into group that Westin refers to as privacy pragmatists. These people make informed cost-benefit decisions regarding Internet privacy (Westin, 2003). However, such reasoning presupposes that respondents accurately assess their level of knowledge and understanding. It is possible that many who believe they are making informed, rational decisions are, in fact, making irrational decisions based on an unrecognized ignorance of the technologies, laws, and data flows related to online information gathering. If people do not understand the technology, regulations, and so forth, then it is unrealistic to expect them to make an informed choice. Such findings could be used to argue for increased government regulation.

Unfortunately there is evidence that many people do not understand the technology, risks, or regulations related to information gathering on the Web. For example, research shows that “many individuals are unaware of the extent of the personal data stored by government and pri-
Ignorance is Bliss

vate corporations” (Roddick, 2001). In addition, the Pew Internet & American Life project found that 56% of Internet users could not identify a cookie and that even those who claimed to have knowledge of cookies seem confused about the technology (Fox, 2000). If people understood the type of information collected, the technology used to gather the information and the potential uses of the information, how would this alter their level of privacy concern and their desire for e-commerce Web site personalization?

In this article, we examine the effect of increased knowledge on privacy concerns and desire for mass customization (Internet shopping preferences). In the following sections, we first review the literature on privacy risks, Web site personalization, also known as mass –customiza-

- tion, and consumer attitudes towards privacy. Next we present a series of hypotheses based on previous research. Then we describe an experiment used to test the hypotheses. Data analysis results are presented. Finally we will discuss the results in the context of earlier research and attempt to explain the implications of our findings.

REVIEW OF THE LITURATURE

Privacy Concerns on the Internet

Since its inception in the mid-1990s, the growth of B2C e-commerce has been phenomenal. According to the Department of Commerce, online retail sales grew from $36.4 billion in 2000 to $54.9 billion in 2003 (http://www.census.gov/mrts/www/current.html). These sales figures could have been many billions of dollars higher except for the fact that many consumers simply refuse to make purchases online. One important factor restraining the growth in Internet sales is the increase in privacy concerns among Internet users (Hoffman et al., 1999). In a survey by AC Nielson, consumers rated the disclosure of personal information and comfort in using a credit card online as the biggest barriers to online purchasing (AC Nielsen, 2001). The Federal Trade Commission estimates that online retail sales were reduced by up to $18 billion due to concerns over privacy (Gellman, 2002). Considerable progress has been made in the development of technological mechanisms for secure payment. Unfortunately while these advances have indeed improved online security, they have done little to alleviate privacy concerns.

Privacy concerns are not merely psychological constructs. There is ample evidence that privacy concerns actually alter consumer behavior in a number of negative ways. According to a survey by AT Kearny, 52% of respondents reported abandoning an online purchase transaction due to privacy concerns (Ragnathan & Grandon, 2002). Total avoidance of online shopping, refusal to provide information and abandoning transactions are not the only responses to privacy concerns. Polls show that 30-40% (Hoffman et al., 1999) of Web users provide false information online. Reasons given include the desire to remain anonymous, avoidance of spam e-mail, and concern about how the Web site will use the information. A consequence of this online lying is that much of the information collected by Web sites is wrong. This both increases the cost and decreases the value of the data collected (Gellman, 2002).

Electronic commerce has increased people’s concerns about privacy because the Internet environment allows business to collect and analyze more personal information with greater ease and efficiency than ever before. “Enormous amounts of consumer data have long been available through offline sources such as credit card transactions, phone orders, warranty cards, applications and a host of other traditional methods. What the digital revolution has done is increase the efficiency and effectiveness with which such information can be
collected and put to use” (Adkinson, 2002). Studies have shown that the amount of data collected is truly vast. A March 2000 Federal Trade Commission survey found that 97% of sites collected some form of personal information. However, only 20% implement the fair information practice principles of notice, choice, access, and security (FTC, 2000).

Firms use several methods to collect information about visitors to their sites. These methods include registration forms, Web surveys, order forms, spyware and cookies. The information thus gathered serves as an important input into marketing, advertising, customer service, and product-related decisions made by online retailers. However, the collection of this information creates the risk of possible misuse and generates concerns over information privacy. In a report to Congress the FTC cited a survey showing that 92% of households with Internet access stated that they do not trust online companies to keep their personal information confidential (FTC, 2000).

Smith et al. (1996) suggest several dimensions of concern related to information privacy. Collection is a general concern that large amounts of personally identifiable data are being collected and stored. Unauthorized secondary use (internal) is the concern that information collected for one purpose could be used for another, unauthorized purpose by the same organization. Unauthorized secondary use (external) is the concern that information collected for one purpose could be used for another, unauthorized purpose after disclosure to an external organization. Improper access is the concern that personal data are available to people not properly authorized to view the data. Errors names the concern that the protections against deliberate or accidental errors are not adequate. One concern that Smith et al. listed as tangential to the privacy issue, but which seems relevant in an e-commerce setting is combining data. This is the concern that several seemingly innocuous pieces of information in disparate databases may be combined to create personally identifying information that the user does not wish to disclose.

Even though all of these specific concerns have been identified, most consumers cannot articulate specific threats to privacy but rather speak of a vague feeling that there is too much information about them “out there.” Unstructured interviews with Internet users show that many have a vague fear that unspecified people unknown to them will have access to personal information. Interviews also indicate that many Web users have little understanding of technology. For example, they don’t really know what a cookie is or how it works. There is also evidence of confusion over privacy rights. Many consumers mistakenly believe that they are protected by “the privacy act” (Gellman, 2002) or that privacy statements are mandated by law.

Whether specific or vague, there is evidence that consumer’s concerns about privacy risks associated with e-commerce are justified. In January 2000, the merger of the online advertising company DoubleClick and the database marketing firm Abacus Direct started a federal investigation when it was revealed that the company had compiled profiles of 100,000 online users, without their consent, and intended to sell the information (Kristol, 2001). More recently, the FTC reported 214,905 instances of identity theft in 2003. This represented 42% of all complaints, up from 40% in 2002 (FTC, 2004). Clearly, some threats to privacy and security related to Internet shopping and online information gathering are real.

Web Site Personalization

E-commerce retailers are increasingly utilizing personalized features on Web sites, also known as mass-customization, in order to build relationships with customers and increase the number of purchases made by each customer (Cranor, 2003). Web site personalization has been shown to be popular with customers and effective at increasing sales (Manber et al., 2000; Personalization...
Consortium, 2000, 2001). However, e-commerce Web site personalization adds new risks that can contribute to an increase in privacy concern. Cranor (2003) suggests four dimensions that can be used to differentiate between types of personalization systems.

Explicit vs. Implicit Data Collection

In explicit data collection systems personalization is based on demographics, preferences, ratings or other information explicitly provided by the user. In implicit data collection systems, personalization is based on information inferred about a user from such things as purchase history, browsing habits, or query history.

Duration

Task or session focused personalization systems make recommendations based on actions taken by the user during the current session. Profile-based personalization systems utilize profiles of users that are updated each time the user returns to the site. These systems often use cookies to recognize customers and retrieve their stored profiles. Logins may also be used to identify the customer and retrieve the corresponding profile.

User Involvement

Personalization may be either user initiated or system initiated. Some systems allow the users to customize their own Web page while other sites attempt to customize for every customer whether they request customization or not.

Reliance on Prediction

Prediction-based personalization systems use explicit or implicit ratings to build a profile that is then compared to others. If similar profiles are found, it predicts that they will have similar preferences and makes recommendations accordingly. Content-based personalization systems also provide customization or recommendations. However, they are based only on the actions of the user, not on profiles of others deemed similar (Herlocker, 2001).

Risks Related to Personalization

Personalization systems can create or exacerbate a number of risks to user privacy. Unsolicited marketing and the possibility of a firm selling information to a third party are generic concerns associated with Internet use. However, these risks are heightened by the fact that personalization systems can collect and store detailed profiles containing considerable amounts of information that go beyond the type of single-purchase information collected by less sophisticated sites.

Another privacy risk associated with personalization systems is fear of prediction. Many people are uncomfortable with the concept that computers are able to make predictions about their habits and interests (Cranor, 2003). Interestingly some users fear that the predictions are too accurate and that the system can figure out information about the user that they would normally be unwilling to divulge. On the other hand, some fear that inaccurate predictions may cause the system to draw incorrect and possible damaging conclusions about the users (Zaslow, 2002).

Personalization systems also increase the risk of inadvertently revealing information to others who use the same computer. Persistent cookies stored on the machine can be used for authentication and provide access to the user’s profile. Through this technology, others using the computer could have access to information such as purchasing records, health queries, and so forth. Profiles often include passwords that then might be used to gain unauthorized access to a user’s other accounts (Cranor, 2003).

In addition, highly detailed profile information could be accesses by third parties, hackers, or e-commerce company employees. It is also possible
that profile information could be subpoenaed in court or used by the government in profiling of surveillance activities. The risks from online profiling led the Federal Trade Commission to recommend that Congress enact legislation to ensure consumer privacy online (FTC, 2000).

Consumer Attitudes Towards Privacy

Alan Westin has been surveying consumers about privacy since 1967. In 1995, the Harris-Westin survey segmented the public on consumer privacy issues into three groups. Privacy fundamentalists rejected consumer benefit claims for data uses and sought legal and regulatory protections. The privacy unconcerned group on the other hand were generally willing to supply their personal information to business and government and were largely unconcerned about privacy risks. Between these two extremes were the privacy pragmatists. According to Westin (2003) this group holds a balanced privacy position. They examine the benefits to them or society of the data collection and use. They wanted to know specific privacy risks and how organizations proposed to control those. Armed with this knowledge the privacy pragmatists would then decide whether to trust the organization or seek legal oversight.

Responses to some of the individual questions used to segment the population can be compared over time to show a trend toward increased public concern about privacy issues. Respondents concerned about threats to personal privacy jumped from 64% in 1978 to 84% in 1995 and had risen to 94% by 1999 (Louis Harris & Associates & Westin, 1999). Surveys in 2000–2002 recorded a dramatic shift in public attitudes (Westin, 2003). By 2002, a majority of consumers (56%) felt that most businesses did not handle the personal information they collected in a confidential manner. This number was up from only 34% in 1999. Sixty-two percent responded that they did not believe that existing laws and organizational practices provided a reasonable level of privacy protection, up from 38% in 1999 (Westin, 2003).

Along with the changes in attitude, the segmentation of consumers has also moved in the direction of increased privacy concern. The privacy unconcerned group dropped from 22% in 1999 to only 8% in 2001. Privacy fundamentalists rose from 25% to 34%. Privacy pragmatists remained the majority at 58%. Policymakers have followed the change in opinion by passing hundreds privacy related laws at the state level and several at the national level. There is an ongoing debate as to the best method to protect privacy while at the same time facilitating electronic commerce. The ability of users to make informed decisions regarding Internet privacy is at the center of that debate.

HYPOTHESES

It is clear that privacy risks exist. It is also apparent that some users do not fully understand the technology or specific risks associated with information gathering on the Internet. This ignorance may explain why users claim that privacy is important but continue to perform activities that put their privacy at risk. In order to investigate if an increase in understanding would alter users’ level of privacy concern or their preferences for Web site personalization, we developed two sets of hypothesis.

Pre-Demonstration Hypotheses

Based on the results of previous research, we tested for between group differences in privacy concern and preferences for Web site personalization prior to the experimental treatment. An Annenberg Public Policy Center survey by Turow & Nir (2000) found that teenagers are more likely to give away personal information than their parents. However, a study by Garbarino & Strahilevitz (2002) indicated that age did not
Ignorance is Bliss

have a significant effect on perceived risk of online shopping. We sought to clarify what effect, if any, that age had on privacy concern or preferences for web-site personalization.

- **H1a**: The level of privacy concern associated with Internet use is affected by age.
- **H1b**: The proportion of users who prefer a Web site to remember them is affected by age.
- **H1c**: The proportion of users who prefer that a Web site make recommendations based on previous visits is affected by age.

Mayer (2002) found difference between Internet users based on education level. As the education level of participants rose they were more likely to say that a promise of confidentiality would make a difference in their decisions, that there were risks associated with a lack of confidentiality, and that there was no guarantee associated with a promise of confidentiality. We developed three hypotheses to test for the effect of education.

- **H2a**: The level of privacy concern associated with Internet use is affected by education level
- **H2b**: The proportion of users who prefer a Web site to remember them will be affected by education level.
- **H2c**: The proportion of users who prefer that a Web site make recommendations based on previous visits is affected by education level.

According to Fernandez & Miyazaki (2001), those who use the Internet frequently have a low perceived risk about the associated technologies. Those who purchase online infrequently will consequently have less knowledge of online technologies, and may be shocked by their capabilities. In 2001, Fernandez & Miyazaki found that “higher Internet experience...[is] related to lower levels of perceived risk toward online shopping, which in turn results in higher online purchase rates.” They also state that “experience gained through simple usage of the Internet for non-purchase purposes such as information gathering and noncommercial communication will lead consumers to discover that privacy and security risks are often exaggerated.” Three hypotheses were developed to test for the impact of Internet usage levels.

- **H3a**: The level of privacy concern associated with Internet use is affected by the level of Internet usage.
- **H3b**: The proportion of users who prefer a Web site to remember them will be affected by the level of Internet usage.
- **H3c**: The proportion of users who prefer that a Web site make recommendations based on previous visits is affected by the level of Internet usage.

Multiple studies have shown that trust is a significant antecedent of customer’s willingness to transact with an online vendor (Gefen, 2000; Jarvenpaa et al., 2000). Jarvenpaa et al. posit that the mechanism by which trust affects one’s willingness to transact is by reducing the perceived risk of doing business with the vendor. Perceived risk is a person’s belief in the likelihood that they will be harmed as a consequence of taking a particular action. Gefen (2000) found that an increase in customer trust resulted in a decrease in the perceived risk associated with doing business with an online vendor. The two most widely cited risks associated with online shopping are privacy risk and security risk. Given the demonstrated relationship between trust and overall perceived risk, it is logical to assume that a similar relationship holds between trust and the level of privacy concern. If this relationship does not hold, we might need to re-evaluate the way that we operationalize the construct of privacy concern.
H4: Those who are more trusting of people in general will tend to have lower levels of privacy concern.

Post-Demonstration Hypotheses

The post-demonstration hypotheses are the focus of this article. They are designed to investigate the impact of the experimental treatment, education concerning information gathering on the Internet through a demonstration, on the dependent variables of privacy concern and two questions related to Internet shopping/personalization preferences. The hypotheses also focus investigation on the possibility of differential impact across various groups.

H5a: The demonstration of privacy risks associated with Internet cookies will increase the level of privacy concern.

H5b: The demonstration will have a greater impact on the privacy concerns of older voters.

H5c: The demonstration will have a greater impact on the privacy concerns of those with lower levels of education.

H5d: The demonstration will have a greater impact on the privacy concerns of those who demonstrate lower levels of Internet usage.

H6: The demonstration of privacy risks will decrease the proportion of people preferring Web sites remember.

H7: The demonstration of privacy risks will decrease the proportion of people preferring to receive recommendations from Web sites.

METHODOLOGY

To test whether or not incongruities could be based upon lack of knowledge, we asked respondents a series of questions to measure their level of trust, online privacy concern and Internet shopping preferences. Basic demographics were also collected.

After a respondent answered questions as described above, we attempted to educate them. On one page of an Internet site, we displayed information about that respondent including their IP address, domain name, and computer information.
tion (see Figure 1). The following is the text that accompanied the information:

The information below was collected from you in just a few short moments without your permission. This is only a simple example. Online companies have the capability to collect much more detailed data. Each time you visit a Web site in fact, you are leaving a 'virtual footprint.' Your information may be actively collected and aggregated by third party companies that can then re-sell your profile.

After the respondents viewed the collected information, they were asked to re-answer the questions pertaining to online privacy concerns and preferences for mass customization. In order to ensure participant privacy, the capture of data was kept confidential. Information stored was limited only to responses and did not include any identifying information.

**Measures**

Several demographic questions were included so that respondents could be categorized by age, education, level of Internet usage and general level of trust. The study utilized two dependent variables, privacy concern and preferences for Web site personalization. The level of privacy concern was measured using a two-item, Likert-type scale. The scale demonstrated a reliability coefficient of .861. Two binary response (Yes/No) questions were used to capture the respondent’s preferences toward e-commerce Web site personalization.

1. Do you prefer for Web sites you visit to remember you?  
2. When you visit a site, do you prefer recommendations based upon your previous visits?

**Data Collection**

The investigation was conducted via an Internet survey. The survey instrument was initially reviewed by three experienced industry professionals to help ensure content validity. A Web site containing the survey was developed. We utilized a convenience sample. The sample included employees from five different companies in the southeast United States as well as undergraduate and graduate students from a southeastern university. Of the 124 useful responses collected 70% were employees of various regional firms and only 30% were students.

**Data Analysis**

Several different analytical techniques were used to analyze the data. Simple T-tests were used to test between-group differences in privacy concern. Z-tests of proportion were used to test for differences in the proportion of respondents answering the binary response format (yes/no) questions related to Internet personalization preferences. ANOVA (SPSS General Linear Model) was used to test for differences between pre-test and post-test scores for privacy concern and for interaction effects. In order to test for differences in the proportions related to the binary response questions from before and after the treatment, McNemar’s test for significance of changes in bivariate nominal scale data was used.

**RESULTS**

**Demographics**

In analyzing our sample, we found that most of the respondents were heavy Internet users. In fact, 54% claimed to spend 10 or more hours per week on the Internet and the majority, 83%, had
purchased more than one item online during the last year. Furthermore, respondents were also mainly college graduates with 72% having a bachelor’s degree or above.

**Age**

In order to determine if age did in fact affect respondent’s reactions to our questions, we broke our sample into two age groups – 18 to 30 and 30 and above. Results indicate no significant difference in the level of privacy concern between the two groups (Table 1). Analysis of results utilizing the Z test of proportion indicate no significant difference between the age groups related to preferences for Internet shopping site personalization (Table 2). These results are consistent with the findings of Garbarino & Strahilevitz (2002) that indicated that age did not have a significant effect on perceived risk of online shopping.

**Education Level**

Respondents were divided into two groups based on education level. One group included those with a bachelor’s degree or higher and the other group include those with less than a bachelor’s degree. Results indicate no difference in privacy concern between the two groups (Table 3). These findings are inconsistent with the findings of Mayer (2002), who found difference between Internet users based on education level. Mayer reported that as the education level of participants rose they were more likely to say that there were risks.

---

**Table 1. H1a: The effect of age on privacy concern**

<table>
<thead>
<tr>
<th>Age</th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>t</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-30</td>
<td>74</td>
<td>3.06</td>
<td>.653</td>
<td>1.042</td>
<td>.299</td>
</tr>
<tr>
<td>30+</td>
<td>50</td>
<td>3.19</td>
<td>.652</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2. H1b&H1c: The effect of age on Internet shopping personalization preferences**

<table>
<thead>
<tr>
<th></th>
<th>18-30 (n=74)</th>
<th>&gt;30</th>
<th>Z</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1b: Prefer Web sites to remember them</td>
<td>Yes 39.2%</td>
<td>40.0%</td>
<td>0.089</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>No 60.8%</td>
<td>60.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1c: Prefer recommendations based on previous visits</td>
<td>Yes 35.1%</td>
<td>42.0%</td>
<td>0.772</td>
<td>.44</td>
</tr>
<tr>
<td></td>
<td>No 64.9%</td>
<td>58.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ignorance is Bliss

Table 3. H2a: The effect of education level on privacy concern

<table>
<thead>
<tr>
<th>Education Level</th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>t</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assoc. Degree or less</td>
<td>35</td>
<td>3.114</td>
<td>.718</td>
<td>-0.014</td>
<td>.989</td>
</tr>
<tr>
<td>Bachelor’s or Higher</td>
<td>89</td>
<td>3.112</td>
<td>.664</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. H2b&H2c: The effect of education level on Internet shopping preferences

<table>
<thead>
<tr>
<th></th>
<th>Assoc. or less (n=35)</th>
<th>Bachelor or higher (n=89)</th>
<th>Z</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1b: Prefer Web sites to remember them</td>
<td>Yes 37.1%</td>
<td>40.4%</td>
<td>0.337</td>
<td>.74</td>
</tr>
<tr>
<td></td>
<td>No 62.9%</td>
<td>59.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1c: Prefer recommendations based on previous visits</td>
<td>Yes 22.9%</td>
<td>43.8%</td>
<td>2.38</td>
<td>.017*</td>
</tr>
<tr>
<td></td>
<td>No 77.1%</td>
<td>56.2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

associated with a lack of confidentiality. Although there was no difference between the two groups in the proportion that preferred Web sites remember them, there was a difference in the proportion that preferred recommendations from the Web sites. 43.8% of those with a bachelor’s degree or higher preferred recommendations based on previous visits compared to only 22.9% of those with less than a bachelor’s degree. This difference is significant at the .05 level (see Table 4).

Internet Usage

Our sample was split into two groups—those online more than 10 hours per week (heavy Internet user) and those online less than 10 hours per week. Results indicate a significant difference (p = .011) between the two groups in their level of privacy concern. Interestingly, the heavy Internet users displayed higher levels of privacy concern than the less frequent users (see Table 5). There were no significant differences in preferences for personalization between the two groups (Table 6). The findings related to level of privacy concern may be inconsistent with the findings of Fernandez & Miyazaki (2001). According to Fernandez & Miyazaki (2001), those who use the Internet frequently have a low perceived risk about the associated technologies. The difference in findings may be accounted for by the slightly different de-
Ignorance is Bliss

Fernandez & Miyazaki (2001) used “perceived risk” as opposed to our construct of “privacy concern.” Given that perceived risk is a more general construct that includes both privacy concerns and security concerns it may be possible for experienced users to have a higher level of privacy concern while still perceiving a lower level of overall risk than less experienced users. While little has been done to address privacy risks, security technology, encryption, e-wallets, and so forth have made financial transactions on the Net much more secure over the last few years. It is possible that experienced users’ knowledge of security technology and the low risks related to online use of credit cards might account for the lower level of perceived risk. If experience is related to knowledge of the technologies and risks associated with e-commerce then these experienced users may be more aware of the privacy risks associated with persistent cookies, data miners, spyware and other information gathering technology. If true, this would be consistent with our findings. Likewise, the lower level of privacy concern reported by those who used the Internet less supports the idea that ignorance is bliss. Those with the lowest level of use, and presumably a lower level of understanding, report the lowest level of privacy concern.

Table 5. H3a: The effect of Internet usage levels on privacy concern

<table>
<thead>
<tr>
<th>Internet Usage Level</th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>T</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;10 hours/week</td>
<td>57</td>
<td>2.947</td>
<td>.698</td>
<td>2.56</td>
<td>.011*</td>
</tr>
<tr>
<td>&gt;10 hours/week</td>
<td>67</td>
<td>3.253</td>
<td>.629</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6. H2b&H2c: The effect of Internet usage levels on Internet shopping preferences

<table>
<thead>
<tr>
<th>H3b: Prefer Web sites to remember them</th>
<th>&lt;10 hrs. week (n=57)</th>
<th>&gt;10 hrs. week (n=67)</th>
<th>Sig. Z</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>43.9%</td>
<td>35.8%</td>
<td>0.92</td>
<td>.36</td>
</tr>
<tr>
<td>No</td>
<td>56.1%</td>
<td>64.2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H3c: Prefer recommendations based on previous visits</th>
<th>&lt;10 hrs. week (n=57)</th>
<th>&gt;10 hrs. week (n=67)</th>
<th>Sig. Z</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>35.1%</td>
<td>40.3%</td>
<td>0.59</td>
<td>.55</td>
</tr>
<tr>
<td>No</td>
<td>64.9%</td>
<td>59.7%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Level of Trust

- **H4:** Those who are more trusting will tend to have lower levels of privacy concern.

To determine if trust affects the level of privacy concern, our sample was split into two groups—high and low trust consumers. Those who answered that they tend to trust people were placed within the high-trust category whereas those who did not tend to trust people were placed in the low-trust category. Results of a t-test indicate that low-trust individuals exhibited higher levels of privacy concern. Since privacy concern is related to a fear of potential harm from others, it is consistent that those with low levels of trust would report higher levels of privacy concern. Given the similarity of the two constructs, this finding could be interpreted as evidence of convergent validity of the privacy concern scale.

**Post-Treatment Results**

Hypothesis 5a represents the main focus of this study. The results of the paired samples t-test indicate a significant difference for all users between pre-test and post-test scores. Analysis shows that educating users about the technology of information gathering resulted in a significant (p < .000) increase in the level of privacy concern (see Table 8 and Figure 2).

ANOVA SPSS (general linear model for repeated measures) indicates a significant main effect of the treatment but no significant interaction effect between treatment and age (see Table 9).

---

**Table 7. Effect of general level of trust on level of privacy concern**

<table>
<thead>
<tr>
<th>General Level of Trust</th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>T</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>25</td>
<td>3.340</td>
<td>.607</td>
<td></td>
<td>.048*</td>
</tr>
<tr>
<td>High</td>
<td>99</td>
<td>3.056</td>
<td>.684</td>
<td>2.04</td>
<td></td>
</tr>
</tbody>
</table>

**Table 8. H5a: The effect of TREATMENT on privacy concern**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>t</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>124</td>
<td>3.112</td>
<td>.677</td>
<td>-6.869</td>
<td>.000**</td>
</tr>
<tr>
<td>Post-test</td>
<td>124</td>
<td>3.419</td>
<td>.679</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 2. The effect of the treatment on all users

Table 9. H5b: The interaction effects between age and TREATMENT on privacy concern

<table>
<thead>
<tr>
<th>Source</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-Subjects Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREATMENT</td>
<td>44.302</td>
<td>0.000**</td>
</tr>
<tr>
<td>TREATMENT x Age_Group</td>
<td>.091</td>
<td>0.763</td>
</tr>
<tr>
<td>Between-Subjects Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Group</td>
<td>0.433</td>
<td>0.785</td>
</tr>
</tbody>
</table>

Table 10. H5c: The interaction effects between TREATMENT and education level on privacy concern

<table>
<thead>
<tr>
<th>Source</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-Subjects Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREATMENT</td>
<td>45.923</td>
<td>0.000**</td>
</tr>
<tr>
<td>TREATMENT x Education Level</td>
<td>1.740</td>
<td>0.190</td>
</tr>
<tr>
<td>Between-Subjects Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education Level</td>
<td>0.282</td>
<td>0.596</td>
</tr>
</tbody>
</table>
Results indicated no interaction effect between the treatment and the user’s education level (see Table 10).

An analysis of the results indicates a differential effect of the treatment based on the level of Internet usage (see Table 11). The treatment had a significantly greater effect on those with lower levels of Internet usage (see Figure 3).

The non-parametric test known as McNemar’s test for significance of changes was used to evaluate the data for hypotheses H6 and H7. The test is used to determine the significance of the proportion of respondents that change their preferences after the treatment.

- **H6**: The demonstration of privacy risks will decrease the proportion of people preferring Web sites remember them.

The contingency table above indicates that 25 people responded yes both before and after the treatment. 24 people responded YES before and changed to NO after treatment, 4 people answered no before and yes afterwards, and 71 people answered no both before and after the treatment (see Table 12). The results indicate that a statis-
tically significant number of users altered their preferences after being educated about information gathering technology online. The direction of change was consistent with the hypothesis. A significant proportion of those who initially preferred to be remembered by Web sites chose NOT to be remembered after the demonstration.

- **H7:** The demonstration of privacy risks will decrease the proportion of people preferring to receive recommendations from Web sites.

  The results shown in Table 13 indicate that a statistically significant number of users altered their preferences regarding receiving recommendations from Web sites based on previous visits after being educated about information gathering technology online. The direction of change was consistent with hypothesis H7. A significant proportion of those who initially preferred to receive recommendations from Web sites chose NOT to receive those recommendations after the demonstration.

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>0</td>
</tr>
</tbody>
</table>

*N = 124, Chi Square = 11.135, sig. <0.000***

**DISCUSSION AND CONCLUSIONS**

**Demographic Differences**

Table 14 presents an overview of the pre-demonstration results. Several hypotheses were tested for between-group differences prior to the demonstration. There was no difference in the level of privacy concern between groups based on age or education level. Although others have found significant differences, our failure to discover many differences in the level of privacy concern based on age or education level is consistent with the findings of Ackerman et al. (1999). The demographic variable that did result in a difference between groups was the level of Internet usage, with heavy users (>10 hours/week) displaying a higher level of privacy concern.

Preferences toward e-commerce Web site personalization were also relatively invariant across groups. The proportion of users that preferred for Web sites to remember them was invariant across age, education, and Internet usage levels. The only significant difference was for those who preferred that Web sites make recommendations based on the results of previous visits between groups divided by educational level. Those with
a bachelor’s degree or higher were more likely to prefer the recommendations than those with less education.

The overall paucity of between-group differences based on age and education level simplify the interpretation of the results and improve the generalizability of the post-demonstration findings. It is not age or education level but how much the respondent uses the Internet that has the greatest impact on the level of privacy concern.

Post-Demonstration Differences

The main focus of this research can be found in the pre-test, post-test results (see Table 15). Notice that the direct effect of the treatment was significant on all three dependent variables, (see hypotheses 5a, 6, and 7 in Table 15). An increased awareness of the types of information collected, information gathering technology, and the potential uses of that information significantly increased the level of privacy concerns of users. Such increased awareness

Table 14. Overview of results—pre-demonstration hypotheses

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1a: The level of privacy concern associated with Internet use is affected by age.</td>
<td>Rejected</td>
</tr>
<tr>
<td>H1b: The proportion of users who prefer a Web site to remember them is affected by age.</td>
<td>Rejected</td>
</tr>
<tr>
<td>H1c: The proportion of users who prefer that a Web site make recommendations based on previous visits is affected by age.</td>
<td>Rejected</td>
</tr>
<tr>
<td>H2a: The level of privacy concern associated with Internet use is affected by education level</td>
<td>Rejected</td>
</tr>
<tr>
<td>H2b: The proportion of users who prefer a Web site to remember them will be affected by education level.</td>
<td>Rejected</td>
</tr>
<tr>
<td>H2c: The proportion of users who prefer that a Web site make recommendations based on previous visits is affected by education level.</td>
<td>Accepted</td>
</tr>
<tr>
<td>H3a: The level of privacy concern associated with Internet use is affected by the level of Internet usage.</td>
<td>Accepted</td>
</tr>
<tr>
<td>H3b: The proportion of users who prefer a Web site to remember them will be affected by the level of Internet usage.</td>
<td>Rejected</td>
</tr>
<tr>
<td>H3c: The proportion of users who prefer that a Web site make recommendations based on previous visits is affected by the level of Internet usage.</td>
<td>Rejected</td>
</tr>
<tr>
<td>H4: Those who are more trusting of people in general will tend to have lower levels of privacy concern.</td>
<td>Accepted</td>
</tr>
</tbody>
</table>
also significantly reduced the proportion of users who prefer e-commerce Web site personalization. It significantly reduced both the proportion that wished for Web sites to remember them and the proportion that preferred recommendations from the Web sites based on previous visits.

When it comes to the level of privacy concern, it appears that ignorance is bliss. It is clear from these results that many users do not understand the amount or types of information that are collected about them on the Internet. It also suggests that many have little understanding of the technologies, such as persistent cookies, that are used to collect personal information. It is this pre-existing lack of understanding that makes the effect of the treatment in this experiment so significant in its impact on the level of privacy concern and preferences for Web site personalization. It is logical to assume (although not proven by this experiment) that those who are heavy Internet users would have a better understanding of information gathering technology such as cookies than those who use the Internet much less. Even before the treatment, the heavy Internet users demonstrated higher levels of privacy concern than the light users. Those who used the Web the least had the lowest levels of privacy concern (see Figure 3). Although the treatment increased the level of privacy concern for both groups, the effect was significantly greater for the light users than for the heavy users. The two groups began with significantly different levels of privacy concern. However, after the treatment, both groups showed increased levels but the two groups were no longer significantly different from each other. Those with the least experience seemed to have been the most shocked by the information in the treatment and therefore they made a significantly greater change in their perceptions.

**CONCLUSION**

These findings have implication for consumers, e-commerce firms, and policy makers. On the consumer side, it is essential to protect one’s own
Ignorance is Bliss

privacy. Our results indicate that consumers do not in fact fully understand online data collection and the related privacy issues. Therefore, consumers should begin by educating themselves about Internet privacy risks and technology. Those consumers who are concerned about online privacy should be meticulous in reading privacy policies in order to establish what information is collected and how it will be employed. If a consumer is uncomfortable with the information that will be collected and/or how that information will be employed, they have the opportunity to visit another site or eschew the risky transaction. Consumers should also be aware of the technology available to help protect privacy. Browsers can be configured to block offending cookies. Sites such as bug-me-not.com can be used for anonymous registration. Encryption technology can also be used to help protect personal privacy. In addition, consumers must become aware that an inherent trade-off exists between the convenience of mass customization and privacy risks online. Consumers must weigh the potential benefits versus the potential risks and make an informed choice regarding preferences for Web site personalization.

These finding also have implications for e-commerce firms. Our findings showed increased levels of privacy concern as consumers were made aware of information gathering technology. Olivero & Lunt (2004) found that an increase in privacy risk awareness reduces the level of trust and increases the demand for control over information use, thus complicating the relationship between the retailer and consumer. Other studies have shown that trust is related to the willingness to transact business with an online firm (Jarvenpaa et al., 2000). Firms should not become complacent and assume that consumers trust them and choose to participate in surreptitious information exchange simply because consumers fail to block their cookies. Our findings indicate that it is ignorance, not acquiescence that causes people to allow cookies. If knowledge of Internet information gathering increases, consumers are likely to lose confidence in the online marketplace. Businesses must understand consumers’ concern about these issues and aim to build consumer trust.

At a minimum, an online company should attempt to instill trust in consumers through contracts, such as privacy policies and/or disclosure agreements, that adhere to the FTC’s fair information privacy guidelines of notice, choice, access and security (FTC, 2000).

Policy makers will also have to deal with the issue of online privacy in the coming years. According to the Center for Democracy & Technology (2004), there were at least twelve bills introduced in the 108th U.S. Congress related to privacy and the Internet. We find the number of people that claim to be concerned about privacy is increasing (Turow, 2003); yet people continue to perform activities that are risky in terms of privacy. There is a debate as to the best method to protect privacy while at the same time facilitating electronic commerce. Some researchers and interest groups argue for self-regulation on the grounds that consumers are free to make a choice with respect to participating in online information exchange (Turow, 2003). Their argument is bolstered by the findings of Westin (2002), who submits that the majority of Internet consumers are “privacy pragmatists.” His description of these users suggests that they make informed cost-benefit decisions regarding Internet privacy risks. Our findings dispute this assertion. The results of this study suggest that a significant number of people do not understand the technology and risks associated with surreptitious information collection on the Internet. If people do not understand the fast-changing technology, regulations, and privacy laws that govern Internet privacy, it is unrealistic to expect them to make an informed choice. Our findings support the need for further regulation and increased consumer education concerning surreptitious data collection on the Internet.
REFERENCES


Turow, J., & Lilach, N. (2000). *The Internet and the family 2000: The view from parents, the view from kids*. Annenberg Public Policy Center of the University of Pennsylvania.


*This work was previously published in the International Journal of Information Security and Privacy, edited by H. R. Nemati, Volume 1, Issue 2, pp. 74-92, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).*
Chapter 7.12
The Function of Representation in a “Smart Home Context”

Mats Edenius
Stockholm School of Economics, Sweden

ABSTRACT
In our society, we seem almost completely engaged in a variety of representational processes. By rendering events and processes “still”, they can be more easily manipulated and transferred into a stock of movable resources which can provide the possibility to control. It is in that context information technology can give us power. In this article it is argued that in order to understand information technology we must begin with representation. I will illustrate such a perspective by a case study that puts a smart home-technology in focus and exemplifies how it may let us deeper our knowledge about IT, how that knowledge is constructed, what actors are involved, what drives them and what kinds of issues are at stake.

Antiquity had been a civilization of spectacle. “To render accessible to a multitude of men the inspection of a small number of objects”: this was the problem to which the architecture of temples, theatres and circuses responded. With spectacle, there was a predominance of public life, the intensity of festivals, sensual proximity. In these rituals in which blood flowed, society found new vigour and formed for a moment a single great body. The modern age poses the opposite problem: “To procure for a small number, or even for a single individual, the instantaneous view of a great multitude”. (Foucault, 1977, p. 216)

INTRODUCTION
We utilize different tools to represent the world and relevant information and knowledge in order to run our lives and make strategic choices. In our contemporary life, we seem almost completely engaged in different representational, fixation — “still” making — processes. A three-dimensional world is reduced to a two-dimensional representation, a lot of reproduced events and objects in a curtailed form (cf. Cooper, 1992; Latour, 1987; Edenius & Borgerson, 2004; Bloomfield & Vurdubakis, 1997; Kallinikos, 1996).

By rendering events, processes and happenings
The Function of Representation in a “Smart Home Context”

“still”, their manipulation and transference into a stock of movable resources can be facilitated. Thus understood, representation provides the possibility to maintain control by expected reactions to the changes made in and through the representations. By representation, I mean a symbolic codification, an operative scheme; “something for something else” (Castoriadis, 1987) like texts, documents, models, or statistics. A commander of a battlefield, for example, needs a map (representation) of the ground to lead his troops, a manager needs an organization chart to manage the company, we create budgets, Internet-home-banks, timetables, and so forth. Variant technologies of representations exist in almost every facet of our lives, not the least of which are our homes. We not only use different representational practices in our homes that are linked to videos, cellular phones, cameras, and so on, but also for cooking facilities, home computers, e-mailing, heating switches, laundry booking, and more.

During the last 10 years or so, a variety of technologies have come into our homes. Many of these could be said to be an application of the intelligent building concept of residential property. When diverse technologies of representations are consolidated in one place in our homes, we use the epithet “smart homes”. Smart (home) services technology was originally developed for the automation of commercial buildings, and has subsequently gained widespread acceptance (Peterson, Williams, & Mills, 2001). Chapman and McCartney (2002) stress that the “smart home” is comprised of everything from “an intelligent building that provides a comfortable and productive environment through automated control systems such as fire safety, security and energy/lighting management” to claims about the “smart home project that allows individuals using it in their daily lives to use their intelligence and think and act for themselves” (Stelcner, 1998; Woodnutt, 1998) to different unobtrusive monitoring systems that support people (Bowes & McCollan, forthcoming; Fisk, forthcoming).

In conventional research about the smart home, a number of epistemological questions are put forward about the world we have attempted to capture. A few of these questions are: Is it well described? What details have been missed? Do we have the necessary features that are required to run the home efficiently? Have we found an acceptable solution for people with restricted mobility? In what way has the smart home improved the user’s life? What kinds of activities do we need to monitor in order to successfully meet our objectives (Graham-Rowe, 2004; Glascock & Kutzik, 2005; Fisk, forthcoming; Kinder, 2000)? Beneath the surface of these studies, we can see structurational perspective posit technology as embodying structures (which may become appropriated by the users during their use of the system, Orlikowski, 2000). The residents’ actions play a crucial role in explaining the consequences associated with the use of new information technology. However, the conventional studies in the field of smart homes leave the analysis on a more or less instrumental level. A more structurational perspective, which is inherently dynamic and grounded in continuous action, is hitherto lacking or is barely discernable.

I will argue that by not examining the epistemological questions posed in the conventional research, there is a potential risk of comprehensively overlooking where the technologies used in a smart home may lead and what the overall results will be. In other words, I would like discuss what a smart home is as much as what it does.

This article is written in the spirit of Sotto (1997) when he, in his ambition to explore what is inherent in information technology, argues that it “can only be undertaken from the point of its ‘essence’ not that of its use. Only then is it possible to perceive its intrinsic qualities and what modes of actions it provides. An adequate investigation of this cannot be carried out, of
course, by a single researcher. It requires the effort of many and familiarization with a quite new viewpoint and its attendant vocabulary”. This paper’s aim is to contribute towards achieving Sotto’s objective.

A smart home could be said to be the outcome of the function of representation, in other words, “something for something else”, and the opportunity for close packing and the reduction of size and mass. This paper will highlight some specific features of the concept of representation and thereby discuss a phenomenon by illuminating parts of its “essence”. My intention is no just to be theoretical, so I will relate different features of representation and discuss how various codifying processes work in practice. I will do this by illustrating my argumentation with a case study.

Let me return to the subject of what a smart home system has to do with representation. The answer is, to a large extent, that a smart home is built based on charts. These charts have their own instrumentally linked questions; do we have the figures we need in order to represent the world? Are we using the proper key figures that are relevant to running our flats? Do we have organized pictures that correspond to our reality? We are confronted with many different choices for how to represent the world so that the representations enable us to run our homes in an efficient way. It is therefore no surprise that the main purpose of employing these technologies is to make it easier for us to maintain control and order (Chapman & McCartney, 2002).

On the basis of a broader historical evaluation of technology (Mumford, 1934; Winner, 1993) it is possible to assert — in general terms — that the “home” involvement of IT accomplishes much more than the sheer transferring of information and knowledge. According to Heidegger (1977), representation could be linked to modernity and the emergence of the industrial world. Heidegger argues that representation differs from previous modes of knowing in that it is not simply concerned with the duplication of the symbolic coding of the world in its diversity, but rather with the selective objectification of things, states and processes (see also Kallinkos, 1996, p. 118). We can say that in the context of the smart home, representation produces a particular brand of reality and functions as a kind of “world-making” (Goodman, 1978). This world is continually working to structure our thoughts and our thinking and acting processes.

By saying this, I am following the line of thinking in which representation is a matter of intervening and reshaping the social by means. Representation needs to be understood not only as the embodiment of a particular technique or sets of techniques but also “...as a world view whose overall orientation permeates all techniques and defines the social and intellectual territory upon which the very modern notion of the technical can emerge” (Kallinikos 1996, 2004).

What the conventional instrumental discussion about the smart home seems to underestimate, is how the smart home in action works as a dynamic configuration of forces and finalities that also produces knowledge. A smart home intervenes and reshapes knowledge via diverse modes of representation. The “smart home”, instead of being a container-like tool where knowledge is imagined to reside as a kind of stable entity or stock of fixed information, becomes a complex system of discursive practices. A smart home, with all its different discursive practices, can thereby be given an active role, central to instances of new knowledge clusters and new categories of significance (See also Bateson, 1973; Power, 1997; Kallinikos, 1996; Chia, 1996; Bloomfield & Vurdubakis, 1997).

To summarize, one function of representation is to give order and direction and make it easy for the mind and body. From conventional research based on this view, we can regard many applications about smart homes in successful terms, but, at the same time there are further processes going on in the light of the way information is
The Function of Representation in a “Smart Home Context”

represented. These processes seem as important as the first one, but, ironically, their constitutive ordinariness seems to have led to their neglect in the analyses of smart home technologies.

From this background, the overall aim of this chapter can be formulated: to highlight and illustrate how features of representation, with their potential world-making and structuring of our thoughts and actions, can give further meaning to what is occurring when a smart home system is used. My aim is not so much to show that using information technology will generate new knowledge clusters and new categories of significance, rather it is to illustrate how new knowledge is constructed, what actors are involved, what drives them, and what kind of issues are at stake.

The remainder of the paper is organized as follows: the next section discusses the method being used and the case from which the illustrations are generated. This is followed by various illustrations of how a smart home is used in daily life situations. The last part brings together the various claims of the chapter and ends with concluding remarks.

CASE ILLUSTRATION AND METHOD

The empirical material in this article is not intended to be viewed as an endeavor to conceive of a clear order from different observations, like a “pure” induction or a complete “case study”. Verifying clear theories or hypotheses in not the primary goal. The method is therefore to be regarded as abductive (see Hanson, 1958; Eco and Sebock, 1983).

The abductive method encapsulates the kind of uncertainty that is present in my research strategy. Based on an abductive mode of thinking, I assert that my inquiries about the research subject are a “reasonable hypothesis”. I am aware that the validity of my supposition — the relevance of features of representation in this matter — could be called into question. However, I am following the vein of Pierce (1935-1966, 2.663 f) (see also Tsoukas, 1989) who say that making a reasonable guess is the only way of getting closer to attaining new and fruitful knowledge. I will argue that the proposed features of representation, with their potential for world-making and the structuring of our thoughts and actions, can give further meaning to what the smart home is as well as provide us with an understanding of what it does.

By saying that I would like to indicate that in this paper the main ambition is the search for fruitful knowledge and richness of good points, rather than trying to reach “full” reliability in my conclusions. Such a standpoint is not odd in social science (see for example Alvesson and Sköldberg (2000) in general social science and Charniak and McDermott (1985) in computer science) — but is not so frequently made explicit in texts.

I will base my argumentation on a variety of illustrations, with the ambition to demonstrate that features of representation are adequate to use in the context of smart homes. The empirical material is based on a building that has used a smart home system for almost three years. There are 59 flats in total and they range in size from 46 to 181 square meters. The material and furnishings are of a rather high standard. The first occupants moved into the building in February 2002.

In Sweden, there is a special type of residential housing where one pays a lump sum and acquires membership in a housing association which owns the building. The right to occupy a specified flat comes with this membership. The daily running of the building is controlled by a board that consists of elected members of the housing association.

The smart home solution encompasses energy use, lighting and heating control, weather forecast, home-security, reception boxes, a family calendar, e-notes, e-mail, “security lock”, the booking of common facilities, lighting control and heating control, and more. The whole system is controlled by a screen/portal in every flat.

I know this building very well because my
family and I (my wife and two children) moved into a 147-meter flat in the building in Spring 2002. One part of the flat has a separate exit door and bathroom, so it is actually two flats. My illustrations are based on my firsthand experiences of living in a smart home building.

The illustrations are not generated by a robust completed Action research project -Action research is much more demanding (cf. Argyris, Putnam, & MacLain, 1985). My research is grounded on a more micro-action level, but it could contribute to scientific research nevertheless. As Gloster (2000) outlines, the Action research method is differentiated between micro Action research, which improves the practical affairs of a particular social system, and Action research, which, in addition, contributes to social scientific research, where all micro-action research can become Action research. The illustrations presented below are based on my own experiences and on a number of discussions between our neighbors. Despite this, I cannot ignore my preunderstanding and my background in the academic field. This chapter is based on empirical material from my experience of moving into a smart home building. Coming from the academic field, it is natural for me to translate my experience into a theoretical framework.

I will present my illustrations as short stories and short descriptions. I am thus following the line of thinking in which narrative knowledge is a major bearer of knowledge in our society (Lyotard, 1984). A common research method is to follow a logic-scientific type of strategy. Nevertheless, when such a strategy is boiled down to grasp knowledge, we realize that even the most “scientific” paper could not be written without stories about how to write, what scientific knowledge is about, and so forth. Narratives are a common and natural mode of communication; people tell stories when they are together, to convince other people, to learn, and to teach. I will tell short stories which are based on my intentions and actions, but they are not just stories. I will mix objective and subjective aspects of a phenomenon; I will occasionally look for causal connections, and by doing this — in the spirit of scientific knowledge — also suggest general laws — but not prove them scientifically. Such narratologi puts the reader in focus, a reader that is able to see how a text was made, not because she divines the writer’s intentions or comprehends universal human natures, but because reader and writer are both producers and consumers of the same set of human institutions (see for example Czarniawska [1998, 2000] who has brought such a perspective into social science).

**KNOWLEDGE PROCESSES WHEN GAIN DOES NOT COME AGAIN**

Cooper (1993) asserts that information theory begins with the construction of a representation (pattern, picture, model) of some aspect of the world. Information is then defined as that which adds to a representation; it is a gain because it tells you something you did not previously know. The function of a gain (or advantage) is echoed in the prefix of the representations. Representation is simply a device for realizing gain and ensuring that the gain comes again (cf. a timetable). The problem with the portal and information technology in our building was that representation did not always occur. The gain did not come again. The information technology was supposed to provide control, but the control was soon quite limited. Those of us who lived in the building had bought a pilot system without knowing it. The technology was very unstable. But stability — in terms of a technique of representations (here the interval by which representations occur) — was not just about having a screen and different functions that switched off occasionally: its effects were more profound in the space of representation.

The occupants of the building soon realized
The Function of Representation in a “Smart Home Context”

that they could not trust the IT solution. People started to talk about how frequently the system was closed down and what to do about it. This discourse became one of the most repeated by the people living in the building. Different discursive formations appeared. One was about what was responsible for causing the problem and what to do about it. Another was about people being angry about what they had bought. It did not stop in terms of discourses; different knowledge processes appeared and they handled the problem.

The occupants began to adapt to the new situation and started to organize their lives according to a smart home that they could not rely on. I will start with one example of my own way of handling an unstable system. We used electronic keys to open the door to the entrance hall and common spaces, and the door to our flat. The locking system was also linked to the smart living concept. We could activate a security system by “secure locking” the door with the electronic key. The key could be used in combination with a personal code and this could activate the security system. This security system included fire, flooding, and burglar alarms. It also locked the door more securely from the inside as well as the outside. When the alarm was activated, it sent a SMS message to a number that could be specified in advance via the terminal.

In the beginning, we trusted the smart living technology. The first summer we lived in the flat we lent a part of it to some friends. What we did not realize was that both doors were activated by the secure lock. When we secure locked our door we also locked the other one. Our friends had to phone us — we were on vacation in Finland — in order to get out of our flat because the door could not be opened from the inside. Our friends threw the key out of a window, and it was picked up by a friendly neighbor who went up and deactivated the system so the door could be opened. From this experience we realized that the system was linked to both doors of the flat. We never used the security locking after that, and I do not know if electronically separating the flats is even possible. What I do know is that because of this glitch we decided to install a separate, analogue (ordinary mechanical) lock into both our front doors.

Another illustration of the same phenomenon could be linked to the reception boxes which could be booked from the portal. The initial test of the system was successful. To get into the main entrance, we simply had to punch in a special pin-code in front of the building. The same code could be used for getting into one of the reception boxes. It is important to note that we could only book the reception boxes when the portal was in steady mode.

A year ago, I accidentally left my keys at home. My wife — who also has a key — was away, and I had to pick my children up at school and take them home for dinner. How was I to get into the flat? Fortunately, I had given a set of spare keys to one of my neighbors in the building. I phoned them and they said they had my keys, but that they would be gone by the time I got home. I thought this was the right time to use the reception boxes, but I was suddenly struck by a feeling of mistrust. I found myself asking my neighbors to give the keys to a shop owner nearby whom I knew quite well. I picked up my keys from the shop owner, felt secure, and the problem was solved.

A third illustration is our possibility to see our consumption of electricity, gas, and water on a daily basis, and to compare our use to the average use in the building. We occasionally tried to use this service in the beginning, but we soon realized that the figures could not be accurate. After a while, we were informed that there were some weaknesses in the hardware, and that the problem would be solved. Despite this, we never used this function again.

We can certainly speculate about why many of the services in the smart home were never used again. We can link this phenomenon to cognitive capacities, interests, and an ocean of human weaknesses. Just the realization that we did not use it again is enough for me. We could
argue that it is important that a smart home is “smart” from the beginning. We could say that smartness is a function of representation, and when representation does not give clarity and direction we seem to adapt to the new situation in a quite pragmatic manner. I am not claiming to be an extraordinarily pragmatic person, but I have been confronted with an unstable system. We are reversibly stuck.

Another significant aspect of the smart home was how we became organized thanks to the possibility to see (much more clearly) what was happening in the building. Of course I am talking about the effects of the opportunity of close packing, the reduction of size and mass, and the ease with which we can see much more clearly. Here we see the importance of representation again.

To see and to see much more clearly. To represent the world in an abbreviated mode is not an innocent project. To be able to see “everything” not only makes it easy for the mind and body, but it provides the possibility to rule/manage the world and make things happen, and generates a controlling mechanism. Undoubtedly, the visual has been dominant in modern Western culture in a variety of ways. Maybe the most quoted author in this field is Foucault (1973, 1977, 1980). His works are largely concerned with representation, vision, and control. He has illuminated the role of the gaze in institutions such as the School and the hospital. Foucault’s (1977) attempts to establish the link between how we represent things and their effects find their highest expression in the notion of Jeremy Bentham’s Panopticon.

The Panopticon consists of a large courtyard with a tower in the center and a set of buildings, divided into levels and cells, on the periphery. In each cell, there are two windows: one brings in light and the other faces the tower, where large observatory windows allow for the surveillance of the cells. The cells are like “small theatres in which each actor is alone, perfectly individualized and constantly visible” (Foucault 1977, p. 201). Rather than depriving the prisoner of light, enclosing and causing physical pain, the Panopticon fully exposes the prisoner. By using this method, disciplinary effects are not achieved through threats and the execution of physical torture and punishment; rather, discipline is achieved through the prisoners themselves; the “prisoner is totally seen, without ever seeing” (Foucault 1977, p. 202). Power, or its corollary control, is distributed. It does not reside within a locus; for “…he who is subjected to a field of visibility, and who knows it, assumes responsibility upon himself; he inscribes in himself the power relation in which he simultaneously plays both roles: he becomes the principle of his own subjection” (Foucault 1977, p. 201). The Panopticon as a form of visibility has led the prisoner to internalize the external constraints associated with control; such constraints now operate from the inside of the body in order to control it and shape it in prescribed ways. Thus there is no need for arms, physical violence, material constraints. Just a gaze. An inspecting gaze, a gaze which each individual under its weight will end up interiorizing to the point that he is his own overseer, each individual thus exercising this surveillance over, and against, himself. Traditional forms of control pale in comparison to this superb formula which is more cost-efficient and operates from a distance.

The panoptic schema, without disappearing as such or losing any of its properties, was destined to spread throughout the social body; its vocation was to become a generalized function…has a role of amplification, although it arranges power; although it is intended to make economic and more effective, it does so not for power itself, nor for the immediate salvation of a threatened society: its aim is to strengthen the social forces — to increase production, to develop the economy, spread education, raise the level of public morality; to increase and multiply. (Foucault, 1977, pp. 205-206)

What Foucault is really describing here — al-
though in metaphorical terms — is the common and well articulated reason for implementing a smart home. The panopticon is also the kind of formalization of a smart home that makes structures clear. The whole neighborhood is potentially represented, visible and transparent. The residents were put into an architectural and administrative space in order to apply classes, numbers and names. The architecture of the smart home system also served the discourse and a kind of formal organization (cf. Cooper, 1993). Division created vision.

Likewise, offices and factories have always been designed with the efficiency of the gaze in mind. We can see it in Taylor’s idea in Scientific Management and Hawthorne’s experiments of how design affects behavior. This ambition to see and control is often embedded in the architectural design of places (Duffy, 1997). It lies at the heart of the ordering of space, tools, and people. It is a gaze that arrests the flux of phenomena, which in turn makes manipulation as well as surveillance possible.

With this background in mind, I will tell one story about the lack of the gaze and its results. Thereafter, I will return to what occurred when we began to use the system to see (much more clearly) what was taking place. By doing this, it was possible to identify different phenomena.

When they built and programmed the smart living system, they decided to keep the people who used the booking functions anonymous. The reason for doing this was to maintain the integrity of everyone living in the building. Again, we can see how a social term, a social epithet, is closely connected to representation.

After just a few weeks of moving into the flat my — at that time — four year old daughter forgot her teddy bear in the laundry room. Unfortunately, our laundry time had run out and we could not get into the laundry room. Who was the person who had booked the laundry room after us? We did not know. The system had backfired. In older houses that have a “smart living” system it is possible to book a time in the laundry room by writing your name on a sheet of paper on the door of the laundry room. By using this type of information technology — old-fashioned representational technology — everyone knew who had signed up to use the laundry room. We, who live in the most sublime IT-house, did not have access to this information. We had to wait outside the laundry room door for someone to come down or we had to leave a note on the door with our phone number. Paradoxically to the IT-definition-discussion above, we had to move in order to get information. We could not control our life solely from the portal. Our power was limited. From this perspective, we could say that the remote control was pulling our strings. We had almost no control in this situation. The mobility function materialized into the shape of a physical guard, waiting outside the laundry room. However, we learned our lesson. We found ourselves becoming disciplined, in other words, making sure we did not forget anything in the laundry room. Another way to view this is to say that the potential control mechanism of the smart living system was decentralized by sitting in front of the laundry room waiting for someone to come and open the door.

The possibility of transparency also affected us in other ways. Everyone who lives in a housing association flat — in Sweden anyhow — knows that there is one main source of agitation, and it is the laundry room. The system helped us to organize when each person in the building could do his/her laundry. When you have done your laundry you are expected to clean up and all the cleaning tools are in the laundry room. Unfortunately, not everyone does this cleaning. This is not a big problem in our building because some of the residents do their duty and the laundry room is quite clean, but it is always an issue that creates annoying discussions among the people in the residence.

Knowing who did not clean up after themselves would increase the frequency of the laundry room getting cleaned. I would venture to say that almost
The Function of Representation in a “Smart Home Context”

everyone who lives in the building would prefer a system in which it is possible to know who is occupying the laundry room at any given time. The transparency would most likely start to discipline the residents, and without anyone having to tell their neighbors to do their duty. This type of power would work automatically. The knowledge that someone could locate you would probably be enough to keep you in line (like in a Panopticon). Thanks to the portal, people could be divided and easily inspected, thus enabling classification and counting, which are the rudiments of representation. A large number of people could be represented in the small space on the screen and inspected at first glance.

It is not true that it is impossible to locate people in the smart living system. It is possible to find such data by delving deeper into the IT-system. However, this is only possible for people who have permission to do this. In our case, certain housing association board members have this kind of permission. This was an opportunity we began to take advantage of and appreciate. During the first year, we realized that it was difficult to find an available time slot when booking the laundry room. This was due to the fact that it was overbooked. How come? We used the possibility to locate people and discovered that three families were using the laundry room extremely frequently. One evening some of the board members went around and asked these families why they booked the laundry room so often. No one wanted to have to tell someone else how often they should use the laundry room, but we noticed that just by asking these questions — or by letting them know that they were being watched — that their way of booking the laundry room changed. We also installed a kind of IT-police. We programmed the IT-system to make it impossible to book more than two time intervals in advance.

Another example is the building where there is a parking area. To get into the garage you need an electronic key-card. Every flat-owner has their own card and their own parking spot. After a while we realized that one resident was using the parking area for other vehicles than his/her own. This was not allowed. Some of the board members started to take a closer look at who the owners of the cars were. We confronted the person who had misappropriated the parking spot, with these facts, but without any real results. It was difficult to find clear evidence and it is irritating to have to take on the role of a policeman in your own building. Suddenly, something occurred that turned the whole situation around. We realized that the person who had misused the parking area also started to use a card that belonged to another resident. The “smart living” system provided us with evidence that enabled us to confront the owner of the residence. It is easier to present evidence that is generated from a system than to say you have been spying on someone in a garage. The link to the perpetual gaze in the IT-system made this relationship more abstract, and, paradoxically, easier to present.

A last example of the effects of transparency is a story about how our sauna was used one day in December 2004. One evening a neighbor of ours smelled smoke coming from the sauna/relaxation room. He knocked on the door to the sauna, but nobody responded. He knocked again because he heard voices coming from the sauna, but no one opened the door. My neighbor — as a board member — picked up a special key which was programmed to open all doors belonging to common facilities (one board member has permission to have this type of key). When he opened the door he saw four young boys who were very drunk. The atmosphere was in no way threatening and the young men said that a friend of theirs who lived in the building had booked the sauna, but that he had gone to pick something up. It was potentially easy to locate the flat from which the sauna was booked. Again, we can see how the entire system facilitates the ability to see what is transpiring. The point I would like to emphasize is that the transparency in the system is not always — separate from its ability to inspect — a significant.
thing by itself, but its power is amplified when it is combined with other IT-applications. This was a Friday evening in December. The man who had thrown the four young men out of the sauna sent an e-mail at 10:49 a.m. the day after and told the other board members about what had happened. Just a few minutes after he had sent the e-mail, another board member was able to tell everyone else that they knew who was in charge of booking and letting the young men into the sauna. She could also tell the others that this was not the first time the young men had had a drinking party in the sauna. Just a few hours after this e-mail arrived, another board member sent an e-mail that confirmed the story and stressed the importance of communicating what was going on in the sauna to the owner of the flat (the parents of the young men). At 5:45 p.m. another board member could tell the others that he had contacted the parents, that they were upset and that they promised that this would not happen again. This kind of rapid action was effective.

**CONCLUDING REMARKS**

The article has provided several illustrations where representations have been put into focus. I have furthermore illuminated different knowledge processes that have resulted. It is argued that in order to understand what information technology is about, we have to start with a discussion about its “essence”, rather than its use.

The discussion in this article does not aspire to reach any clear specified closure, but perhaps it will be a starting point for further discussion of the features and functions of representation and thereby the inherent “economy” in smart homes at work. The technique of representations exemplified in the text is, of course, not unique to smart homes. It could also be applied to many other technologies. However, the smart home could be regarded as one technology where the power of representation has been exaggerated and hyperbolized by different techniques of representations. I have stressed how important it might be to regard a smart home as something that is not preformed by pre-existing agents. Instead, I have stressed that users in a smart home context are organized as they participate in a smart home.

Representation is a key to understanding what is happening in this type of process. If we begin to think in these terms instead of in “what happens” terms to generate theories about smart homes and ICT in general, we might better understand how we develop deeply entrenched thought patterns which unnecessarily can circumscribe the possibility for action; and by that the theory could also become instrumentally usable (Chia, 2000). By being aware of the power of representation in smart homes solutions we can organize the way we run our residential housing in better ways. For example, we should be aware of that the possibility of using the system to locate individuals will probably be used. Secondly, and therefore, it may be a good idea to let everybody living in residential housing know about this possibility in advance and that it may be used (or not). This could be done by writing down clear rules about this matter. Another important thing to be aware of is to be sensitive to instability in smart home systems. Instability will create knowledge, and change the routines and the way we use the system, and routines are hard to break.

Hence, one of the essential implications of this study is the significance of broadening the research criteria for studying the smart home. Traditionally, researchers focus on different kinds of interviews, or text-based material that arrests what people do in a smart home context and how the technology could be developed in tune with the users’ wishes. A fuller understanding of the smart home requires one not to overlook the fact that people know what they do and why they do it, nor to suggest that the conventional studies are not productive. Instead, it would perhaps be beneficial to go beyond the conventional parameters that are found in conventional research about smart
homes, by also discussing what effect peoples' actions have. Does the smart home generate trust or mistrust by its controlling mechanisms? People who live in a smart house, do they really want to adapt to new ICT-solutions once the they have been confronted with solutions linked to instability or even worse; when they have been aware of its real benefits in real situations?

If we return to the perspective of regarding information technology and the smart home system as representations of control in one’s life, we are back at Mumford’s (1934) well-known example about the man trying to split the stone, but not into larger pieces that he can carry. The knowledge that the practices surrounding the smart home generate, and what it demands from us seems to be a reminder not of the user’s physical, but rather his or her cognitive limits, or will to control/manage. However, those of us who live in a smart home are not necessarily more rational, cognitively gifted or clumsy than others, or more stuck by managing our life and neighborhood. Our advantage, or occasional disadvantage, is that we have a screen on the wall and the others do not.

ACKNOWLEDGMENT

I am very grateful for the insightful comments made by the reviewers and the Editor on previous versions of the paper. I am also thankful to Jan Wallander’s and Tom Hedelius’ Foundations for Social Science Research.

REFERENCES


Czarniawska, B. (2000). The uses of narrative
The Function of Representation in a “Smart Home Context”

in organization research. Göteborg, Gothenburg Research Institute.


ENDNOTE

1 The smart living system encompasses a lot of different functions. The most important are presented in the appendix.
APPENDIX

<table>
<thead>
<tr>
<th>System/Equipment</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Home network</strong></td>
<td><strong>Energy use</strong>&lt;br&gt;Presents the measured use of electricity, gas and water on a daily basis as well as over a requested period for each energy type. The resident can also compare this usage with the average use in the building.</td>
</tr>
<tr>
<td><strong>Lighting</strong></td>
<td>The residents can choose which sockets should be on/off at different point of time.</td>
</tr>
<tr>
<td><strong>Heating</strong></td>
<td>The residents can create their own temperature profile for every room in the flat.</td>
</tr>
<tr>
<td><strong>Weather forecast</strong></td>
<td>Shows the weather of today and the following day.</td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td>Includes fire, leakage and burglar alarms. If anything is wrong the alarm is set to send a message to the resident by SMS or e-mail.</td>
</tr>
<tr>
<td><strong>Calendar</strong></td>
<td>A calendar that co-ordinates the family’s activities.</td>
</tr>
<tr>
<td><strong>E-notes</strong></td>
<td>The residents can leave e-notes to the rest of the family and create common reminder lists. Then, SMS or e-mail can send the list to a recipient.</td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
<td>Ordinary e-mail.</td>
</tr>
<tr>
<td><strong>“Away lock”</strong></td>
<td>Controls the main water supply, electricity and/or gas to the stove, lamp in the hall, the burglar motion detector, power sockets, ventilation, lighting and heating.*</td>
</tr>
<tr>
<td><strong>Bookings of common facilities</strong></td>
<td>To operate the laundry room or the sauna/guestroom. The residents can also supervise the washing machines in the laundry room and receive a signal via the laptop computer when the washing machines are done.</td>
</tr>
<tr>
<td><strong>Broadband connection</strong></td>
<td>To surf on the Internet.</td>
</tr>
<tr>
<td><strong>Integrated system for computer and telephone</strong></td>
<td>Same socket in a room can be used for both units.</td>
</tr>
<tr>
<td><strong>Electronic key</strong></td>
<td>To open the door to the lobby and general spaces, e.g., the garage and the laundry room. In the building the residents also use the electronic key to open the door to the flat.</td>
</tr>
<tr>
<td><strong>Security camera at the front door</strong></td>
<td>The residents can both see and talk to the visitors via the camera before letting them into the building.</td>
</tr>
<tr>
<td><strong>Reception boxes</strong></td>
<td>A space for reception of ordered goods. The reception boxes could be booked from the terminal.</td>
</tr>
</tbody>
</table>

*When the entrance door is locked, the water is automatically shut off two hours after locking, as well as gas or electricity to stove, sockets above the worktop, lighting, and heating. When the resident comes home the lamp in the hall lights up and the main water supply, the gas, sockets, ventilation, lighting, and heating are switched on automatically. The light in the hall is on until the resident manually turns it off.*

Chapter 7.13
Towards a Meta-Model for Socio-Instrumental Pragmatism

Peter Rittgen
University College of Borås, Sweden

ABSTRACT

We claim that a general conceptual framework for the IS field should provide some kind of common upper-level ontology to describe and explain artifact-mediated social interaction. Such an ontology, socio-instrumental pragmatism (SIP), has been suggested. Our aim is to refine and formalize this ontology by providing a meta-model in the form of a unified modeling language (UML) class diagram. We discuss the implications of such a model as well as its relation to other ontologies. The meta-model is validated by using it in the evaluation of an existing business modeling language.

INTRODUCTION

The rise in the use of information systems (IS) is undeniable, and every day IS become a more important part of organizations. But far from being perfect, the design and implementation of IS in organizations is still a very problematic task that is often fraught with failure (Ågerfalk & Goldkuhl, 2006). There is a need for a better understanding of IS, organizations, and their relation to come up with a framework capable of integrating these two concepts. For the past two decades, theories of communication have been imported into the IS field and the language action perspective (LAP) has been proposed as a way to understand IS and organizations based on communication (Goldkuhl, 1982; Winograd & Flores, 1986). Later on, an ontology to capture the social world was proposed and described in Goldkuhl (2001), Goldkuhl, Röstlinger, and Braf (2001), Goldkuhl (2005), and Goldkuhl and Ågerfalk (2002). This ontology was named “socio-instrumental pragmatism” since it aims at human actions which are supported by instruments and performed within the social world (Goldkuhl,
Towards a Meta-Model for Socio-Instrumental Pragmatism

Socio-instrumental pragmatism (SIP) presents a generic framework which allows for the analysis of the social world. Within this world there are six ontological categories:

1. Humans
2. Human inner worlds
3. Human Actions
4. Signs
5. Artifacts
6. Natural objects

Since SIP was intended as a generic framework which can serve as a base to analyze the social world, it is not aimed exclusively at the IS field. We think that a meta-model based on the SIP ontology but with a focus on the IS field is needed. This meta-model has its foundations in both LAP and SIP and presents a model that will allow us to view organizations and IS together with a focus on actions.

The model consists of the basic categories actions, actors, and objects. In addition to this we also consider other important aspects of organizations that are related to their functioning.

**TOWARDS A META-MODEL SOCIO-INSTRUMENTAL PRAGMATISM**

As mentioned before, there is a need for a framework that allows us to describe social systems in a clearer and more thorough way. Our work is based on the SIP ontology. Within the SIP ontology there are six ontological categories (Goldkuhl, 2002):

- **Humans** are the most important participants in the social world described by the SIP ontology; they act in the world based on

![Figure 1. Realms of the world within the SIP ontology (Goldkuhl, 2002)](image)
meanings and perceptions that they derive from the world.

- **Human inner world** represents the knowledge that a human being has acquired over time about themselves and the external world; this inner world is intended to be seen as part of the human being.

- **Human actions** also form a part of the human being; they can be overt, which means that the actions are intended to intervene in the external world, thus trying to change something about it. And they can be covert when they are aimed to change some human being’s inner world; covert actions try to change knowledge that is present in the human inner world.

- **Signs** are the result of communicative actions; for instance, when write a note saying, “I will be at the store”, the writing of the note is by itself a communicative action but the note created is a sign which will mean something to the person that will read it.

- **Artifacts** are things which are not symbolic and not natural but which are material and artificially created. Examples of artifacts are cars, clothes, a knife, and so forth. The difference between signs and artifacts is that while signs are intended to mean something to someone (symbolic), artifacts perform material actions. For instance, a human might use a knife (artifact) to cut some carrots, that is, artifacts are needed to perform material actions.

- **Natural environment** are the objects present in the environment that are not artificially created by humans (e.g., trees).

Figure 1 shows the different realms of the world according to the SIP ontology.

**META-MODEL**

Our model is divided into three main categories:

- Actions
- Actors
- Objects

Although we do not see Agent as a category; we do acknowledge the importance of agency and describe it as a special element in the model.

**Actors**

Actors are the main entities in our model, and they can perform either as locutor or addressee within the communicative context. When actors perform actions that are directed towards another actor we speak of social actions. They can be performed either in a human-human relation or in a human-artifact-human relation. When performing as locutor the actor is trying to change some aspect of the world by means of his/her actions. For instance, when a person pays the phone bill she is trying to avoid the interruption of her phone service. When performing as addressee the actor receives and interprets an action directed to him and can act himself as a consequence of that action. Taking our example the addressee will be the phone company, which at the moment of receiving the payment will not make any attempt to interrupt the customer’s phone service.

Besides locutor and addressee we can distinguish between organizational actors and human actors. The former is an actor that performs as an agent on behalf of the organization, the latter performs an action on behalf of herself.

**Objects**

An object may be physical or conceptual and it may be formed by other objects or related to
Towards a Meta-Model for Socio-Instrumental Pragmatism

them, but every object is unique (Embley, Kurts, & Woodfield, 1994). Under the object category we have artificial and natural objects. Artificial are those that are created by human beings; natural objects are those created by nature and found in the environment. Among the artificial objects, we have artifacts (material objects) and signs (can be material or immaterial). Artifacts are created to extend actors’ capabilities. An artifact is seen as a tool. Signs on the other hand are not tools but messages in a static phase waiting to be interpreted by actors or artifacts. A message can take either a physical form (a written text) or a non-physical form (an utterance) (Goldkuhl, 2002).

We can distinguish between four different types of artifacts: static, dynamic, automated, and multi-level. Static artifacts are those that cannot perform any operation by themselves, for example, a stone, a knife, or an axe. Dynamic objects are those capable of performing some operations by themselves but they need constant control by a human being to function properly, for example a car or a drill. Automated artifacts are those that can operate entirely by themselves and only need to be started by an actor. Here we can mention a washing machine as an example (Goldkuhl & Ågerfalk, 2005).

Multi-level artifacts are those that have a mix of capabilities and can perform either as static, dynamic, or automated artifacts depending on the circumstances. Multi-level artifacts have an important property which is the capability of creating and interpreting signs. They lack consciousness and are ruled by a predefined set of instructions that serve as a guide to perform the predefined actions they do. IT systems are an example of multilevel artifacts. Signs can be created either by human beings or artifacts, and every sign can be interpreted by human beings only, by artifacts only, or by both (Goldkuhl, 2005). A written note is a sign, an utterance performed by an actor is another example of a sign as well as a ticket printed by a system in an electronic store.

Actions

The objective of human actions is to change something in the world. They can be communicative or material. The main difference between these two types of actions lies in the fact that communicative actions are intended to change knowledge. Knowledge is implicitly meaningful to someone; and knowledge handling is an exclusive characteristic of actors within an IS. On the other hand, material actions are aimed at material conditions and aspects of the world which are meaningful to someone. They are intended to change something physical among the external world. Winograd and Flores (1984) stated that language is prior to consciousness and we might add that consciousness is prior to actions performed by actors. As a human characteristic, knowledge can be learned through actions, either communicative actions (for instance, a conversation) or material actions (e.g., when studying an object). Knowledge is the result of the actor’s interpretation of both communicative and material actions, and it can be acquired in a social context (from other actors transferring knowledge, for example, in a classroom) or in a non-social context (a person reading a book on her own) (Goldkuhl, 2001).

We can divide actions into i-actions (intervening actions) and r-actions (receiving actions). I-actions are those intended to make a change in the external world, for example, the action of opening a window is intended to change a particular aspect of the external world (the window will move from closed to open). R-actions are those executed covertly, for example when two people are going out and person A tells B “It’s cold outside” (communicative i-action). Then person B listens and interprets the message (r-action) and maybe after that person B will take a jacket on the way out (material i-action) (Goldkuhl, 2001). Among i-actions and r-actions we have indefinite and predefined actions.
Towards a Meta-Model for Socio-Instrumental Pragmatism

Indefinite actions are those performed by humans and we call them indefinite since it is not certain how they will be performed by the actor. The same action can vary from actor to actor. When two employees are ordered to clean a shelf, they will both do it but not in the same way; one can do it better or faster than the other one. Indefinite actions can be either r-actions or i-actions. On the other hand we have predefined actions which are performed by artifacts. These actions will always be performed in the same way following previously programmed instructions (Goldkuhl, 2005). Predefined actions are i-actions, since they are intended to change an aspect of the external world. Among indefinite and predefined actions we find both communicative and material actions.

Both types of actions are aimed at changing an aspect of the world surrounding the actor or artifact but we can see communicative actions as a two-phase action (at least) where the actor A performs a communicative action that is intended to change an aspect of the world but is directed toward another actor or artefact B in the first phase. In the second phase B (if A was successful) executes the action that A desired. Although material, the last action can sometimes be performed without an initial communicative action.

Organizational actions can be either internal or external, and material or communicative. Material and communicative actions within organizations form patterns. Although human beings perform the actions within organizations, we can say that an organization can act. An organizational action has human origins and purposes and is done through humans, by humans, or by artifacts that act on behalf of the organization (Goldkuhl et al., 2001). We will consider organizational actions that constitute an interaction of two or more elements from the organization (actors or artifacts) within an organizational context. We can say that a worker at a clothing factory using a sewing machine to manufacture clothes is performing an organizational action. He is acting to perform an organizational objective (to produce clothes). But, for instance, a man on a farm that goes to the forest to chop wood using an axe, although using an artifact to perform the action, is not performing an organizational action since there is no organizational purpose if he merely burns the wood to warm up his house.

When performing actions by means or with the help of IT systems, we can distinguish between three different types of actions: interactive, automatic, and consequential actions. Interactive actions are supported by and performed through IS and they consist of one or more elementary interactions. Elementary interactions (e-actions) consist of three phases: a user action, an IT system action, and a user interpretation (Goldkuhl, 2001). Let us take the example of an online bank transfer done by the user online. The user will initially introduce his username and password to access the bank system (phase 1), after this the IT system will check in the database if the information is correct and if it is it will grant access to the user and display a welcome screen (phase 2). The welcome screen is interpreted, and the user now knows that he can start his transaction. This is the end of the elementary interaction. Later on the user inputs the data to make the bank transfer, such as account number, amount to be transferred, and so forth (phase 1 of a second e-interaction), and so on.

Automatic actions are performed by IT systems that produce messages for the actors or other systems. They are done entirely without human intervention. Let us take the banking system again: after logging on, a message pops up in telling the customer that the due date for the credit card payment is very close. The system will execute this operation by itself and present it to the user.

Consequential actions are those performed as a consequence of a message. Taking the bank example again, when the customer sees that his payment is due he might proceed to execute the
Towards a Meta-Model for Socio-Instrumental Pragmatism

payment, or he might decide not to do it and wait for the final day.

Based on these types of IS actions, IS are seen as information action systems. This perspective is called actability. Actability is supposed to reinforce the concept of usability within the IS framework and focuses on action and communication. IS actability is the information system’s ability to perform actions, and to permit, promote, and facilitate the performance of actions by users, either by means of the system or based on information provided by it in a business context (Sjöström & Goldkuhl, 2002). An IS is said to be actable when it has the following characteristics (Cronholm & Goldkuhl, 2005):

- Clear action repertoire
- Good communication satisfaction
- Easy to navigate interface
- Action transparency
- Clear feedback
- Easy access to action log
- Personalized information
- Familiar vocabulary
- Good support for business actions
- Capability to understand different communicative intentions

The components of the IS are the IT system, the actor, and the e-action. IT systems are social systems that are technically implemented and have an action memory which stores the past actions and some future actions. Actors can play the role of communicator, performer, or interpreter in the IS.

Agents

Agents are a special type of object; we can position agents between objects and actors. They are created by actors, and perform actions to help them complete their tasks. They can be seen as servants of actors, but they have a level of communicative capabilities that allow them to act as communicative mediators, and they are also capable of creating signs for the actors or other agents to interpret. Agents have a transformative capability, a property that human beings have as well. The difference between agents and human beings lies in the fact that human beings can perform both socially aware actions (such as a conversation) and nonsocially aware actions (such as a blink) while agents can only execute the latter (Rose & Jones, 2004).

IT systems can perform as agents, but describing an IT system can be a very tricky task due to the versatility that these artifacts have. An IT system can either be seen as a static artifact (e.g., when we are reading an e-mail), it can be seen as an automated artifact (e.g., a payroll system from a bank that executes the payments for the employees automatically every 15 days), and it can be a dynamic artifact (e.g., a sales system used by a sales person in an electronics store) (Goldkuhl & Ågerfalk, 2005). In all three cases there is a common denominator: communication. In the first case the IT system is acting as an intermediate device between the sender and the receiver of the e-mail. In the second case it can also be seen as a communicative mediator between the employer and the employee, that is, by executing the payments it communicates to the employees that their employer is paying them. In the third case the system acts also as a communicative device between the salesperson and the customer. The IT system gives information about the products to the salesperson which is transmitted to the customers and it also prints a ticket of the sale which is taken by the customer. This ticket communicates to the customer what she bought and how much it cost. The IT system in this case executes other tasks which may turn into communicative actions. It decreases the inventory of the article sold and if the article inventory is low it will communicate to the inventory manager that the article is running out.

Communication is seen as a kind of action that IT systems can perform and by doing so they
Towards a Meta-Model for Socio-Instrumental Pragmatism

become communication mediators. IT systems as well as actors have the capability to create signs and process them (in the case of the IT system) and to interpret them (in the case of actors) (Goldkuhl, 2001). The relation between the signs and the interpreters/processors of them is called pragmatics. Messages are a product of communication and are also an important prerequisite of it. Within IS pragmatics, actions are divided into those that occur within the sign transfer and consequential actions that are performed in response to the transferred sign (Goldkuhl & Ågerfalk, 2002).

Organizational Actions

Roughly, we can say that within an organization every actor acts to fulfill organizational objectives; hence, they are agents helping to accomplish organizational actions. Let us take the example

Table 1. Sales process for an electronic store

<table>
<thead>
<tr>
<th>Actor</th>
<th>Action</th>
<th>Type of action</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salesperson</td>
<td>Utterance: May I help you?</td>
<td>Communicative i-action</td>
<td>S performs an i-action</td>
</tr>
<tr>
<td>Customer</td>
<td>Utterance: Yes I would like to buy some batteries</td>
<td>R-action; followed by a Communicative i-action</td>
<td>C interprets the utterance performed by S (r-action) and makes an utterance (i-action)</td>
</tr>
<tr>
<td>Salesperson</td>
<td>Utterance: Which type of batteries?</td>
<td>R-action; followed by a Communicative i-action</td>
<td>SP interprets the utterance performed by C (r-action) and makes an utterance (i-action)</td>
</tr>
<tr>
<td>Customer</td>
<td>Utterance: Rechargeable AA batteries please</td>
<td>R-action; followed by a Communicative i-action</td>
<td>C interprets the utterance performed by S (r-action) and makes an utterance (i-action)</td>
</tr>
<tr>
<td>Salesperson</td>
<td>Take the batteries and pass them through the bar code reader</td>
<td>R-action; followed by an Interactive i-action</td>
<td>S interprets the utterance performed by C (r-action) and performs an interactive i-action</td>
</tr>
<tr>
<td>IS (Agent)</td>
<td>Reads bar code and gets the information for the product from database</td>
<td>Automatic action</td>
<td>IT system performs an automatic action and displays articles details on the screen</td>
</tr>
<tr>
<td>Salesperson</td>
<td>Gets information of product and tell it to the customer</td>
<td>R-action; followed by a Consequential i-action</td>
<td>S interprets the message on the screen (r-action) and makes an utterance (i-action), telling the customer about the details</td>
</tr>
<tr>
<td>Customer</td>
<td>Pays for the batteries</td>
<td>R-action; followed by a Material action</td>
<td>C interprets the message of S and performs a material action (payment)</td>
</tr>
<tr>
<td>Salesperson</td>
<td>Receives payment and give receipt</td>
<td>R-action; followed by an Interactive i-action; followed by an Automatic action; Material action</td>
<td>S receives the payment and closes the sale in the IT system, IT system executes the automatic actions of modifying the inventory and printing the receipt, S hands the receipt to C (material action)</td>
</tr>
</tbody>
</table>
Towards a Meta-Model for Socio-Instrumental Pragmatism

of an electronics store. A customer (C) comes into the store and the following dialog with the salesperson (S) develops.

(S): “May I help you?”
(C): “Yes, I would like to buy some batteries.”
(S): “Which type of batteries do you want?”
(C): “Rechargeable AA batteries please.”
(S): “We have X and Y brands.”
(C): “I would like X.”
(S) passes the batteries over the bar code reader and says “$10, please.”
(C) pays.
(S) completes the sale in the system and hands the receipt to the customer.

When we analyze this business interaction according to our meta-model we arrive at the results shown in Table 1. We will see organizational actions as those actions performed to fulfill an organizational objective. In Table 1, this objective is to sell batteries. We can also notice that many of the actions are multi-functional, that is, one “surface” action corresponds to a number of implied, “hidden” actions. When the salesperson asks for the type of batteries, he performs an implicit r-action by correctly interpreting the request “I would like to buy some batteries”. At the same time he also responds appropriately by performing the i-action of asking for the type of batteries.

As a result of the contemplations in the previous sections we have developed a meta-model (see Figure 2) that covers the most important aspects of socio-instrumental pragmatism as discussed in the relevant literature. Technically the meta-model takes the form of a UML class diagram with generalization/specialization and association.
OTHER ONTOLOGIES

In the literature, we can find a number of competing ontologies that are also potential candidates for a general conceptual framework for the IS field. We look at three of them and discuss their role in such a framework and their relation to SIP.

Social Roles

This theory emphasizes the importance of social roles among the social concepts. Social roles are concepts that can be played by certain entities when they interact with other entities. Examples of social roles are money, professor, and president. The premises behind roles are (Masolo et al., 2004):

- Roles are properties and can be predicated of different entities. This means that different entities can play the same role.
- Roles are anti-rigid; this aspect regards the temporal nature of the relation between roles and their players.
- Roles have a relational nature, being properties, and they have different types of dependencies to other roles/entities. Notional, identificational, and definitional dependence are some of them.
- Roles are linked to contexts; roles are described as determined by external factors (context).

But as mentioned by Masolo et al. (2004), social roles theory “makes use of a simplified ontology, and therefore only partially characterizes social entities.” (p. 5). With the social roles framework we can only partially describe social aspects. We can get a good description of social actors (a professor, a president, an employee) and objects, but we cannot get a description for other important aspects of social concepts such as actions. Another point to highlight is the fact that there is no explicit distinction between actors and objects, since they are both seen as endurants with the difference that objects lack intentionality.

Intentional Collectives

Collections are seen as social objects which depend on member entities and concepts. Several types of collections are distinguished: simple collections (e.g., a collection of stamps), organized collections can be conceived as characterized roles played by members of the collection and that relate among them through social objects. For instance, in a collection of senators, although all of the members have the same role (senator), one of them can be the president of the economic commission, another one can be the secretary of the agricultural commission, and yet another one can be the vice-president of the senate.

Collectives are considered to be something more than collections (Bottazzi, Catenacci, Gangemi, & Lehmann, 2006). Collectives are built around the concepts of intentionality, agent, and plan, the latter being the most important concept within collectives. Intentionality can be seen as the feature by which agents are directed to something. An agent is considered to be intentional and it is oriented towards producing results. An agent function is also to conceive plans. By plan we refer to a description that represents an action schema. Another concept described within the framework of this theory is that of a task, which is a course that is mostly used to sequence activities or other processes that are under the control of a planner. Collectives are a collection of agents. In collectives roles are played by agents and they actively participate in plans and roles.

This theory presents a more detailed conception of social concepts based on social roles. We can find some similarities with our perspective:

- Tasks represent actions within our perspective
• Agents represent actors within our perspective
• Plans can be seen as organizational objectives that lead to the execution of organizational actions

**Actor Network Theory**

The actor network theory (ANT) describes a world containing both human and nonhuman entities but at the same time it makes no difference in the importance of the elements within the network. Instead of separating the social from the technical, ANT analyzes the world in a sociotechnical manner by arguing that a merely technical or a merely social relation is not possible. Instead of actors, the term actant is used in ANT which can be used to refer either to human or nonhuman elements within the network. ANT’s objective is not to analyze the nature and features of the entities in the network, but to study the relations that compound the network, the mechanics of power between the elements, and how the network relations are built (Tatnall & Gilding, 1999).

The ANT approach appears as radical, and the main issue with it is the symmetry between humans, non-human objects, and natural objects. Within our model, the distinction between all these elements is very important in order to understand the functioning and communication within organizations.

For the ANT theorists, the elements of the network relay and prolong collective actions, and no element is considered as the source of the actions. Therefore, instead of being actions, they are seen as events. As we can see, events can be mapped to organizational actions within our perspective. Organizational actions are collective actions performed by one or more actors, agents, or artifacts within organizational boundaries and with a common goal. Thus, events are described in the same way as being collective actions that have no source but that are performed by the actants of the network (McLean & Hassard, 2004).

**EVALUATING A BUSINESS MODELING LANGUAGE**

To validate the completeness and correctness of the meta-model, we investigated an existing business modeling language. We have chosen the language of SIMM (situation-adaptable work and information systems modeling method), Goldkuhl (1996), because we have considerable experience with this language in action research projects. The analysis proceeded in the following way. First we have extracted the constituting concepts from the diagrams. The diagrams are: collaboration graph, interaction graph, process graph, and action graph. These concepts can be found in the left column of Table 2.

In the second step, we have related the concepts of SIMM to the categories of SIP as formalized in the meta-model. For this step we have thoroughly analyzed the documentation of SIMM (Röstlinger & Goldkuhl, 2006). The result of this step is presented in the middle column of Table 2. The right column contains a textual description of the respective concept.

In the third step, we have analyzed the table to identify deficiencies of the meta-model and the language SIMM. We found that all business concepts of SIMM could be captured accurately by some category of the meta-model. Only concepts originating from a different domain (knowledge) or belonging to general upper-level ontology (time and place) could not be represented.

On the other hand, we discovered some minor shortcomings of SIMM. One is that of construct (or concept) overload where one concept takes on different meanings in different contexts and is therefore mapped to several ontological categories. An example of that is *Initialization*, which can be both an actor and an object/sign.

Another issue is that of construct (or concept) redundancy where one and the same ontological category can be expressed by several concepts of the modeling language. An example of this are *directly related actions* and *Indirectly related actions*. 

---

2399
### Table 2. Mapping SIMM concepts to meta-model categories

<table>
<thead>
<tr>
<th>SIMM concept</th>
<th>Related SIP concept</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity/action</td>
<td>Action</td>
<td>Describes what is done and by whom (executor); this is complemented by place, time, and instrument of execution</td>
</tr>
<tr>
<td>Actor</td>
<td>Actor</td>
<td>Person, group of persons, role, organizational unit</td>
</tr>
<tr>
<td>Alternative actions</td>
<td>Action (with conditional sub-actions, # = 1)</td>
<td>One of two or more actions/action sequences is carried out</td>
</tr>
<tr>
<td>Alternative prerequisites/results</td>
<td>Object (with conditional sub-objects, # = 1)</td>
<td>One of two or more objects is required for or the result of an action</td>
</tr>
<tr>
<td>Artifact</td>
<td>Artifact</td>
<td>Artificial system that performs actions automatically, for example, an IT system</td>
</tr>
<tr>
<td>Being</td>
<td>Actor/object</td>
<td>Human being or animal</td>
</tr>
<tr>
<td>Cancellation</td>
<td>Object/sign</td>
<td>The action is canceled when a certain object is present or at a certain time</td>
</tr>
<tr>
<td>Closed information</td>
<td>Sign</td>
<td>Information that can only be interpreted with the help of some instrument</td>
</tr>
<tr>
<td>Collaboration object</td>
<td>Object</td>
<td>Information, material</td>
</tr>
<tr>
<td>Combined actions</td>
<td>Action (with concurrently ordered sub-actions)</td>
<td>Two or more actions/action sequences are carried out</td>
</tr>
<tr>
<td>Combined or alternative actions</td>
<td>Action (with conditional sub-actions, # &gt; 0)</td>
<td>One or more of two or more actions/action sequences is carried out</td>
</tr>
<tr>
<td>Combined or alternative prerequisites/results</td>
<td>Object (with conditional sub-objects, # &gt; 0)</td>
<td>Two or more objects are required for or the result of an action</td>
</tr>
<tr>
<td>Combined prerequisites/results</td>
<td>Object (with sub-objects)</td>
<td>Two or more objects are required for or the result of an action</td>
</tr>
<tr>
<td>Composite action</td>
<td>Action</td>
<td>A number of actions that is carried out together</td>
</tr>
<tr>
<td>Composite executor/unit</td>
<td>Actor/artifact</td>
<td>Several executors that together form a named unit for execution</td>
</tr>
<tr>
<td>Composite process</td>
<td>Action</td>
<td>A number of processes that are combined to a meaningful unit</td>
</tr>
<tr>
<td>Condition</td>
<td>Sign</td>
<td>Condition for prerequisite, result or action</td>
</tr>
<tr>
<td>Conditional action</td>
<td>Action (conditional)</td>
<td>Action that is carried out or not depending on some condition</td>
</tr>
<tr>
<td>Customer</td>
<td>Actor</td>
<td>A special role of actor as the final receiver of a product</td>
</tr>
<tr>
<td>Data storage</td>
<td>Artifact</td>
<td>Place to store closed information</td>
</tr>
<tr>
<td>Directly related actions</td>
<td>Action (with sequentially ordered sub-actions)</td>
<td>Actions that are immediately follow each other</td>
</tr>
<tr>
<td>End</td>
<td>Sign</td>
<td>End of process</td>
</tr>
<tr>
<td>Executor</td>
<td>Actor/artifact</td>
<td>Performer of an interactive action</td>
</tr>
<tr>
<td>Executor</td>
<td>Actor/artifact</td>
<td>Performer of an action in a process</td>
</tr>
<tr>
<td>Executor</td>
<td>Actor/artifact</td>
<td>Person, group of persons, role, organizational unit or artifact</td>
</tr>
<tr>
<td>Independency relation</td>
<td>Action (without effect)</td>
<td>Control flow between actions</td>
</tr>
<tr>
<td>Independent actions</td>
<td>Action (with sequentially ordered sub-actions)</td>
<td>Sequential actions that have no causal relation</td>
</tr>
</tbody>
</table>

*continued on following page*
Table 2. Mapping SIMM concepts to meta-model categories (continued)

<table>
<thead>
<tr>
<th>SIMM concept</th>
<th>Related SIP concept</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirectly related actions</td>
<td>Action (with sequentially ordered sub-actions)</td>
<td>Actions that follow each other but not immediately</td>
</tr>
<tr>
<td>Information</td>
<td>Sign</td>
<td>Messages that can be interpreted directly by human beings</td>
</tr>
<tr>
<td>Information relation</td>
<td>Communicative action</td>
<td>Flow of information between executors/units</td>
</tr>
<tr>
<td>Information relation</td>
<td>Communicative action</td>
<td>Flow of information between actions</td>
</tr>
<tr>
<td>Information storage</td>
<td>Artifact</td>
<td>Place to store information/messages</td>
</tr>
<tr>
<td>Initialization</td>
<td>Actor</td>
<td>Actor that starts a mutually interactive action</td>
</tr>
<tr>
<td>Initialization</td>
<td>Object/sign</td>
<td>An action is started when a certain prerequisite becomes available or at a certain time</td>
</tr>
<tr>
<td>Instrument</td>
<td>Object/artifact</td>
<td>A tool that is used in an action</td>
</tr>
<tr>
<td>Interaction partner</td>
<td>Actor/Artifact</td>
<td>Executor/unit that performs interactive actions</td>
</tr>
<tr>
<td>Interaction sequence</td>
<td>(Ordered) actions</td>
<td>A number of actions that is ordered in time</td>
</tr>
<tr>
<td>Interactive action</td>
<td>Action</td>
<td>An action that the sender directs towards the receiver</td>
</tr>
<tr>
<td>Knowledge</td>
<td>(refers to an ontology of the mind)</td>
<td>Internal state of the mind</td>
</tr>
<tr>
<td>Marked actor</td>
<td>Actor</td>
<td>A focused actor</td>
</tr>
<tr>
<td>Marked collaboration object</td>
<td>Object</td>
<td>A focused collaboration object</td>
</tr>
<tr>
<td>Marked interaction partner</td>
<td>Actor/artifact</td>
<td>A focused interaction partner</td>
</tr>
<tr>
<td>Marked process</td>
<td>Action</td>
<td>A focused process</td>
</tr>
<tr>
<td>Material</td>
<td>Object</td>
<td>Material objects that can be accompanied by information</td>
</tr>
<tr>
<td>Material relation</td>
<td>Material action</td>
<td>Flow of material between executors/units or between actions</td>
</tr>
<tr>
<td>Material storage</td>
<td>Artifact</td>
<td>Place to store material that can be accompanied by information</td>
</tr>
<tr>
<td>Mutually interactive action</td>
<td>Action</td>
<td>Two or more interactive actions that follow an initiative/response pattern</td>
</tr>
<tr>
<td>Non-object</td>
<td>Object</td>
<td>Hypothetical object</td>
</tr>
<tr>
<td>Parallel actions</td>
<td>Action (with concurrently ordered sub-actions)</td>
<td>Actions that are performed concurrently</td>
</tr>
<tr>
<td>Place</td>
<td>(refers to an upper-level ontology)</td>
<td>Describes where an action is performed</td>
</tr>
<tr>
<td>Prerequisite</td>
<td>Object</td>
<td>Action object that is required for an action</td>
</tr>
<tr>
<td>Process</td>
<td>Action</td>
<td>A number of actions with a common goal</td>
</tr>
<tr>
<td>Process object</td>
<td>Object</td>
<td>Product/result of a process that is useful for the customer or order that initiates a process</td>
</tr>
<tr>
<td>Producer</td>
<td>Actor/artifact</td>
<td>Executor/unit that produces and directs information and/or material for/towards the receiver</td>
</tr>
</tbody>
</table>

*continued on following page*
Table 2. Mapping SIMM concepts to meta-model categories (continued)

<table>
<thead>
<tr>
<th>SIMM concept</th>
<th>Related SIP concept</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Receiver</td>
<td>Actor/artifact</td>
<td>Executor/unit that receives and uses information and/or material</td>
</tr>
<tr>
<td>Result</td>
<td>Object</td>
<td>Product of the executed activity/action</td>
</tr>
<tr>
<td>Reused process</td>
<td>Action</td>
<td>A process that occurs in different places in the process graph</td>
</tr>
<tr>
<td>Sequential actions</td>
<td>Action (with sequentially ordered sub-actions)</td>
<td>Actions that follow one after the other</td>
</tr>
<tr>
<td>Start</td>
<td>Sign</td>
<td>Start of process</td>
</tr>
<tr>
<td>Time</td>
<td>(refers to an upper-level ontology)</td>
<td>Point in time for execution of an action</td>
</tr>
<tr>
<td>Variants</td>
<td>Action (with conditional sub-actions)</td>
<td>One action is chosen from a number of alternatives</td>
</tr>
</tbody>
</table>

actions. Both concepts cannot be distinguished ontologically which can be seen as a case of over-specification from the point of view of the ontology. In the latter it is not considered to be relevant whether actions follow each other immediately or not.

CONCLUSION

We started our chapter with the assumption that socio-instrumental pragmatism might contribute towards the development of a general conceptual framework for the IS field. We then set out to capture both the breadth and depth of the SIP literature with a suitable meta-model. This process consisted of uncovering the central concepts and their (often implicit) relations and making them explicit in a clear and concise way. We did this with the help of a UML class diagram, a modeling language which is well established and documented and can therefore be expected to support the communication of and about the meta-model among a large group of IS researchers.

We are well aware of the fact that such a meta-model is not, and cannot be, the ultimate solution to a general conceptual framework for the IS field. But we think that it can stimulate a fruitful discussion about the vital components of such a framework. We do not know of any other meta-models that cover the breadth of IT-mediated social action with similar stringency.

REFERENCES


This work was previously published in Managing Worldwide Operations and Communications with Information Technology, edited by M. Khosrow-Pour, pp. 1013-1017, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 7.14
The Effect of Choice and Announcement Duration on the Estimation of Telephone Hold Time

Philip Kortum
Rice University, USA

Randolph G. Bias
The University of Texas at Austin, USA

Benjamin A. Knott
U. S. Air Force Research Laboratory, USA

Robert G. Bushey
AT&T Laboratories, USA

ABSTRACT

If a caller is placed on hold when they call a business, about half will hang up before the call is answered. Of those that hang up, only half of those will call back (Staino, 1994). Optimizing the on-hold experience has the potential to reduce hang-ups and make being put on hold more palatable to the caller. The current study assessed the influence of the opportunity to make a music choice and the length of pre-music announcement duration on perceived on-hold durations and customer satisfaction. Subjective assessments of on-hold times were significantly shorter with longer announcements, but satisfaction did not change. The chance to choose music improved satisfaction, but did not significantly reduce subjective time estimates. To test if multiple within-subject trials might have led to prospective time judgments, a between-subjects design replication was conducted, with each participant estimating only one time interval. A similar pattern of results was obtained.
INTRODUCTION

Many companies provide one or more customer support phone numbers for customers to call when they wish to get information, solicit technical assistance, lodge a complaint, or otherwise make contact with the company. The call to customer support is one contributor to the consumer’s overall user experience, and research shows (Tax, Brown, & Chandrashekaran, 1998), is a major factor in brand loyalty.

The provision of good customer support is therefore an important concern of companies. Along with the quality of information received in the call (defined roughly as the ability of that information to lead to a successful resolution of the consumer’s problem or completion of the consumer’s task), one key variable influencing the consumer’s judgment of the call is the time spent on hold. One way to minimize the consumers’ time spent in the hold queue is to oversupply the call center with customer support representatives. Of course, this is costly, and will ultimately influence the cost of the services and products passed along to the consumer. Therefore, companies attempt to optimize the size of the customer support force, in hopes of keeping the force maximally utilized, while at the same time minimizing the cost to the consumer, in terms of time spent on hold.

The correlation between the length of a time segment and a person’s estimate of the length of that segment is less than +1.0. Put another way, human beings’ perception of the length of time segments is not determined solely by the length of that segment. Rather, many other variables influence the perceived length of time, including variables in the stimuli that fill the time span, variables in the context of that time span, and variables in the person who is perceiving the time span (Block, 1990).

Because time is a limited resource, most people view it as valuable (Feldman & Hornik, 1981). Considering this view, a lengthy hold time may result in customer dissatisfaction regardless of the level of service a customer ultimately receives. There is considerable interest in the telecommunications industry in identifying methods to reduce caller dissatisfaction due to long hold queues for consumer call centers. Telephony and software experts are employed to maximize the efficiency of the call center applications and minimize the average actual time on hold. For times when it may not be possible for a call center to reduce the actual length of time that consumers are placed on hold, it is still possible to change the content presented in the hold time interval so as to reduce the perceived length of the hold time. Two experiments are reported here that advance the understanding of: 1) the role of the caller’s ability to choose content while in the hold queue; and 2) the length of the automated announcement before being placed on hold on the person’s judgment of the length of time they are on hold.

Related Research in Cognitive Psychology

The topic of time perception is one that has been of interest to experimental psychologists and others interested in human information processing for a long time. Researchers have shown that, ceteris paribus, time spans that are filled with content (such as music) or a task to occupy a person’s attention are judged to be shorter than time spans with nothing filling the time (Predebon, 1996). Research on time perception suggests that an individual’s estimate of the duration of a time interval is altered by various attributes of stimuli presented in a given interval. Several studies have shown that the apparent duration of a time interval decreases as the cognitive processing demands increase (Brown, 1985; Chastain & Ferraro, 1997; Fortin & Rousseau, 1987; Fortin, Bourque, & Kirouac, 1993). Qualities of that content that have been shown to influence the perceived time duration include the amount of information memorized during the time period (Zakay, 1989), the complexity of the stimulus (e.g.,
The Effect of Choice and Announcement Duration of the Estimation of Telephone Hold Time

Macar, 1996), the tempo or periodicity of the stimulus (Macar, 1996; Polkosky & Lewis, 2002), the hierarchical nature of the tasks performed (Tractinsky & Meyer, 2001), how interesting passages to be read are (Gray, Gray, & Loehlin, 1975; Hawkins & Tedford, 1975), and the size of the visual display on which information is presented (Bobko, Bobko, & Davis, 1986). Liu and Wickens (1994) found that “Time estimation appeared to be sensitive to the presence of perceptual/cognitive demands, but not to response related activities to which behavioral automaticity has developed” (p. 1853).

Participant variables shown to affect duration estimation include gender (Vercruyssen & Rodenburg, 1992), body temperature (Hancock, 1993), the person’s affective state (e.g., Fiedler, 1988; Hornik, 1993; Thayer & Schiff, 1975), and whether or not the participants know that they will be asked later to estimate the duration of the time segment (Zakay, 1993). Even context of the time interval being estimated, such as time of day (Pfaff, 1968) and length of an immediately preceding time interval (Sasaki, Seutomi, Nakajima, & ten Hoopen, 2002) have been shown to influence duration estimations.

The effect of cognitive complexity on time estimation has been explained in terms of demands of certain stimuli on attentional resources. For example, individuals may track the passage of time with an internal counter and as more attentional resources are allocated to nontemporal stimuli over a time interval, fewer resources may be available for the internal timing mechanism (Thomas & Weaver, 1975; Zakay, 1989). Thus, lower duration estimates result as more attention is allocated to external stimuli.

Various models have been proposed to account for the variations in time segment estimation. Ornstein (1970) proposed a storage size model in which apparent duration depends on the size of the storage space in memory required for storing information about an interval. The model predicts that the more events in a particular time interval, or the greater their complexity, the longer the interval seems.

Attentional allocation models (e.g., Hicks, Miller, Gaes, & Bierman, 1977; Zakay, 1993) maintain that time estimates are based on the output of some sort of cognitive timing device. If a person is engaged in a task that has high attentional demands, the attention that can be allocated to the “timer” is limited, and the interval seems shorter.

A third model, the change/segmentation model (e.g., Block, 1990; Fraisse, 1963; Poynter & Homa, 1983), offers that the number of discrete changes in an interval determine the perceived duration of that interval. Events, that may be external or internal (mental events), segment a time interval, and so durations are judged according to the number of segments, with more events thus leading to a longer estimated duration.

Relevant Applied Research

Beyond the ken of cognitive theory, the topic of time perception has immediacy for designers of call centers, developers of software install programs, amusement park managers, and any number of other practitioners; there are many situations in the business world where it is beneficial for a business entity to reduce the amount of time a customer perceives during the wait for a particular service. The reduction in perceived wait time is not the end goal itself, of course. The reductions in perceived wait time typically result in higher satisfaction ratings from the customer, and higher satisfaction ratings can translate into a number of tangible, positive results for the business. With clear incentives to reduce perceived customer wait time, the goal is to alter the environment and thus the customer experience in such a way that the customers walk away from the transaction happier than they otherwise would because they perceive that the time they had to wait for their service was smaller than it actually was.
Although it might seem appropriate to focus solely on the measure of customer satisfaction as a proxy for reduced perceived wait time (when the assumption is made that lower perceived wait times result in higher customer satisfaction ratings), customer satisfaction ratings alone do not account for all of the dimensions customers use when they are making an assessment of a service. For example, even if a person reported having a magnificent time at Disneyland, that person is also likely to report (and retain the perception) that the wait times for some of the rides tended to be very long. This perception will likely be factored into future decisions about whether or not to visit the park again. As such, both customer satisfaction and perceived wait time must be measured when making assessments of the goodness of a service for which the customer must wait.

While the practice of the manipulation of time perception is evidenced across a wide range of activities where consumers must wait for a product or service, of primary interest here is the ability to minimize the perceived length of time while callers are on telephone hold, and there is ample evidence that this can be accomplished. Sold on Hold, a company that sells advertising systems for on-hold interactive voice response (IVR) systems, conducted a study in which 30,000 callers were placed on hold for 1 minute and then asked how long they had been holding. Three conditions were used: silence, music, and messages. Their data showed that all of the callers exposed to silence estimated that they had been on hold for over 1 minute (with 27% estimating they had been on hold for over 5 minutes). The music and message conditions showed 28% and 17% of the customers overestimating their wait time, respectively (Sold on Hold, 2004). Importantly, the amount of overestimation changed based on what the listeners heard during that wait. It would seem, intuitively, that users might underestimate time more if they are enjoying the music or message that is heard during the wait interval, thus supporting the old adage that “time flies when you are having fun.” Kellaris and Kent (1992) showed that this is not necessarily the case. In their study they presented two kinds of music to users who were on hold: positively-valenced music and atonal music. In contrast to the conventional wisdom, participants exposed to the atonal music underestimated the amount of time on hold more than those callers who were exposed to the positively-valenced (normal) music.

Sometimes the perception of time manifests itself as the time people are willing to wait for a service before they abandon the telephone queue and hang up. In one such study, different types of music were played to people who were on hold for a protective services abuse hotline. Each type of music was used as the hold music for 2 different weeks in a 10-week study. The authors then performed an analysis to see if there were differences in the number of calls that were abandoned (people who hung up before an attendant answered), based on the type of hold music that was played. Five types of music were used: jazz, country, classical, popular, and relaxation. There were significant differences in the number of abandoned calls for the various types of music, with relaxation music resulting in over double the number of abandonments as jazz (Ramos, 1993). Speech-based IVR systems appear to yield similar results. In an experimental situation, telephone users were asked to get a speech recognizer to route them to a particular section of a telephone service. This was a ruse, and no participants were ever routed. During the wait intervals, half the participants heard silence and the other half heard music. At the end of the experiment (5 minutes) the participants were asked how long they had been on the telephone trying to get the system to work. Those who heard music overestimated the time by 5%, while those who heard no music overestimated by 39%. Further, when asked how much longer they would have waited before they hung up, callers who had music indicated that they would have waited, on average, about 20% longer than the users who heard silence.
The Effect of Choice and Announcement Duration of the Estimation of Telephone Hold Time

(Gueguen & Jacob, 2002). In a study conducted by North, Hargreaves and McKendrick (1999), callers responding to an ad in a newspaper for research volunteers were placed on hold with either recorded messages, original Beatles music, or renditions of Beatles music made with pan pipes. The results showed that music that listeners said they liked led to the longest times before they would abandon the queue. Their findings are in opposition to the findings of Kellaris and Kent (1992), suggesting that the mechanisms controlling time estimation and call abandonment, while related, are different.

Of course, the stimulus presented to the user during the telephone wait does not have to be music or messages. IBM used the rate of change of an auditory cue (a ticking sound) to give users information about the response time of a speech-enabled telephone application. They found that increasing the ticking rate of the auditory stimulus increased users’ perceptions of the system’s response time (as well as the user’s feelings of anxiety, stress and impatience). The slower ticking rates produced underestimations of the length of the system response time, but it made the users unhappy (Polkosky & Lewis, 2002). Kortum, Peres, Knott and Bushey (2005) found a more complicated relationship between nonmusical auditory cues and perceived wait time, where the arrangement of the stimulus tones acted differentially on participants’ perception of time. Like Polkosky and Lewis (2002), Kortum et al. (2005) also found that users were generally unsatisfied with the stimuli.

The real cost in dollars of waiting can also impact users’ perception of time. In a simulated banking situation an experimenter had users put on hold while waiting to conduct a financial transaction. Half of the participants were told that the call was to a 900 number (where the caller must pay a per-minute charge) and that their compensation for doing the study would be reduced by the amount of the per-minute charges associated with the 900 call. For short waiting conditions (30 seconds and under), both the free and toll call participants overestimated perceived wait time by nearly 100%. As the wait time increased, overestimation was greatly reduced, but the wait time in the toll condition was overestimated by about 14% more than in the no toll condition (Antonides, Verhoef, & Aalst, 2002).

As all of these studies collectively show, there is ample evidence to indicate that the perception of time can be altered based on the environment to which the user is exposed. The key is to find variables that: a) change the perception of time in the desired direction; b) change the user satisfaction in the desired direction; c) are robust across a wide range of populations; and d) have effects that are large enough to justify the trouble and cost of implementing them.

Motivation for the Current Studies

A study of hold queue content demonstrated the positive impact of implementing an Interactive Voice Portal (IVP) within a call center hold queue (Knott, Pasquale, Joseph, Miller, & Mills, 2003). Specifically, the authors found that when callers used an IVP while waiting on hold, their satisfaction with the wait time was high and they made relatively short wait time estimates compared to several other content conditions (i.e., Silence, Music, and Advertisements). Two explanations were offered by the authors for the results. First, the IVP condition was an active condition when compared to the other more passive content conditions. Research has shown that individuals who are cognitively active underestimate time, compared to less active individuals who tend to overestimate time (e.g., Burnside, 1971; Troutwine & O’Neil, 1981). Second, the IVP shifted the locus of control from the system to the user. First described by Rotter (1966) and then later by Schneiderman (1987) specifically for human-computer interaction, the locus of control describes how much control users feel they have over the system and its operation. An internal locus of control means...
that users believe that their actions have direct impact on the function of the system, whereas an external locus of control means that users believe that input provided does not directly influence system operation. The IVP gave callers a choice over the hold queue content, enhancing an internal locus of control. An internal locus of control can result in a more positive emotional state, leading to higher customer satisfaction. Research has shown that emotional or affective state influences perceived time duration (Hornik, 1993) and the acceptability of a wait time (Chebat & Genlinas-Chebat, 1995). A third explanation for the advantage of the IVP is possible. The IVP included a relatively lengthy hold queue announcement. The hold queue announcement is a brief message that informs the callers that they are being placed on hold. In recalling the amount of wait time, callers may have discounted the time during which the announcement was being played.

The study described here attempts to answer the questions raised in Knott et al. (2003) by determining if the users’ ability to select on-hold music (internal locus of control) and the length of the opening announcement can positively affect both the users’ perception of time and their subsequent satisfaction ratings.

**METHODS: EXPERIMENT 1**

**Participants**

Twenty-four individuals, 13 females and 11 males, were paid (50 USD) to participate in a 1½ session. The mean age was 37, and ages ranged from 23-58 years. Participants were recruited from the general population of a major metropolitan area in the south-central United States. The selection criteria required only that the participants had used a phone before. Because being placed on telephone hold is a rather ubiquitous experience in the United States, this criterion was sufficient to garner a representative sample.

**Apparatus**

The study was conducted in a small laboratory room furnished with a couch, and a coffee table with a Western Electric 2500 Series telephone on it. Calls were placed to an IVR system running on a personal computer through a Dialogic Telephone Interface card. The IVR was programmed using VBVoice.

**Procedure**

Participants were given a brief overview of the study upon arrival and were given time to read and sign an informed consent form. Participants were asked to remove their watches and place them in a drawer in the testing room “so they would focus on the study and would not be attending to the time to see when their session would end.” They were told that this was a standard procedure used for all usability studies. To conceal the purpose of the study, the participants were told that they were there to evaluate several automated systems by placing phone calls to the systems and answering a usability questionnaire. They were also told that they might experience some delay in reaching the automated systems because the systems were currently undergoing heavy testing.

Participants were given a testing packet with instructions, 16 task scenarios, and customer satisfaction questionnaires for each call. The tasks are listed in Appendix A. For each call, participants heard a brief hold queue announcement. They were then placed in the hold queue where they listened to music while they waited to be connected to the customer service system. After waiting on hold, participants were connected to an automated system that asked them for the purpose of their call. Participants were instructed to ask the question provided in their task scenario and to write down the answer they received from the automated system. A different task was used for each call and the tasks were presented in random order. After each call, participants were
The Effect of Choice and Announcement Duration of the Estimation of Telephone Hold Time

Figure 1. An illustration of the four experimental conditions. A) Choice/Short announcement, B) Choice/Long announcement, C) No-Choice/Short announcement, D) No-Choice/Long announcement

Table 1. Short and long announcements for the choice and no-choice hold queue conditions

<table>
<thead>
<tr>
<th></th>
<th>Short Announcement</th>
<th>Long Announcement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Choice Condition</strong></td>
<td>&quot;Welcome to SBC. While on hold, please select your music? For Classical, press one; For Country, press two; For Classic Rock, press three; For Jazz, press four.&quot;</td>
<td>&quot;Welcome to the SBC customer service line. Please hold for the next available service representative. You can listen to music while you are on hold. Please select the type of music you would like to listen to: For Classical, press one; For Country, press two; For Classic Rock, press three; For Jazz, press four.&quot;</td>
</tr>
<tr>
<td><strong>No-Choice Condition</strong></td>
<td>&quot;Welcome to the SBC customer service. In order to provide better customer service, all calls are answered in the order taken. Please hold for the next available service representative.&quot;</td>
<td>&quot;Welcome to the SBC customer service line. Your call is very important to us, but all of our agents are busy assisting other customers right now. Please hold to speak with the next available service representative. Because your call will be answered in the order in which it was received you will get the fastest service by staying on this call rather than hanging up and calling back.&quot;</td>
</tr>
<tr>
<td><strong>Duration</strong></td>
<td>11 seconds</td>
<td>21 seconds</td>
</tr>
</tbody>
</table>
asked to estimate the amount of time they had to wait before reaching the customer service system, in minutes and seconds. They then completed a customer satisfaction questionnaire consisting of 11 statements on a five-point scale (Strongly Disagree to Strongly Agree). Eight of the questions were distractors and related to system usability. The remaining relevant questions appear in Appendix B.

The study used a repeated-measures design in which all participants made 16 phone calls. The experimental conditions were presented in random order for wait time (30, 60, 120, or 240 seconds), music choice (music choice or no-choice conditions) and announcement duration (long or short). For calls in which participants could choose the music type, four options were presented: country, classical, classic rock, and jazz. For the no-choice trials, the type of music that was played was automatically selected by the system and was counterbalanced across calls. The actual music selections played for the users within each type of music were identical for both the choice and no-choice conditions.

The four experimental conditions created by the two choice conditions and the two announcement length conditions are represented in Figure 1. The long and short hold queue announcements for the choice and no-choice conditions are shown in Table 1.

Hypothesis

It was hypothesized that:

1. When callers can choose the music type, time estimates will be shorter and customer satisfaction will be higher.
2. Longer announcements will result in shorter time estimates and higher customer satisfaction.

RESULTS: EXPERIMENT 1

Estimate error is defined as the difference between the actual and perceived wait time, such that positive numbers indicate an overestimation of the actual wait time and negative numbers indicate an underestimation of the actual wait time. As can be seen in Table 2, on the average participants overestimated the wait time, but estimate error was smaller for the long announcement condition (M = 1.05 sec) compared to the short announcement condition (M = 13.9 sec).

A 2 × 2 repeated measures ANOVA was conducted to evaluate the effect of music choice and announcement duration on estimate error. There was a main effect for announcement duration on estimate error (F(1,23) = 4.4, p < .05) (see Figure 2). One-sample t-tests were used to compare participants’ estimates with the actual wait times. Estimate error was significantly greater than zero when the short announcement was used, t(191) = 2.37, p = .02. The difference was not significant for the long announcement condition, demonstrating that participants did not overestimate the wait time for this condition. The main effect of music choice on estimate error was not significant (p = .11).

A customer satisfaction measure was created by taking the average score for the three customer satisfaction questions and converting to a percentage score. A 2 × 2 ANOVA was conducted to evaluate the effect of music choice and announcement duration on customer satisfaction scores. The main effect for music choice was significant, (F(1,23) = 19.3, p < .001) with customer satisfaction higher for the choice conditions (M = 75%) compared to the no-choice condition (M = 64%). The effect of announcement duration on customer satisfaction was not significant, F(1,23) = 3.1, p = .09 (M = 71% and M = 67% for the long and short announcements, respectively). Customer satisfaction is significantly correlated with estimation error (r(382) = -.35, p < .01). Table 3 and
The Effect of Choice and Announcement Duration of the Estimation of Telephone Hold Time

Figure 3 show the customer satisfaction data for experiment 1.

After each trial participants were asked how much they liked the music selection (the actual piece of music played during the hold interval) they had just heard, on a 5-point scale. It was thought that the degree to which people liked the specific music selection might be important, so a Pearson correlation was conducted between Statement 4 (‘I liked the music selection I heard.’) and Estimate Error or Customer Satisfaction scores. The more people liked the music selection played, the lower

Table 2. Estimate error data, in seconds, for experiment one. Negative numbers indicate underestimates.

<table>
<thead>
<tr>
<th>Wait Time</th>
<th>Choice</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Mean</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Short</td>
<td>-1.9</td>
<td>15.9</td>
<td>-5.0</td>
<td>15.0</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>9.7</td>
<td>14.7</td>
<td>3.3</td>
<td>14.7</td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>240</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>Short</td>
<td>4.9</td>
<td>-2.1</td>
<td>2.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Choice</td>
<td>Short</td>
<td>5.7</td>
<td>14.7</td>
<td>3.3</td>
<td>15.6</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>11.7</td>
<td>17.6</td>
<td>17.6</td>
<td>14.7</td>
</tr>
<tr>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>240</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>Short</td>
<td>22.8</td>
<td>4.2</td>
<td>7.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column means</td>
<td>Short</td>
<td>13.85</td>
<td>1.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grand mean</td>
<td>Short</td>
<td>4.92</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Experiment one, estimate error, in seconds, by condition

![Chart showing estimate error by condition]
The Effect of Choice and Announcement Duration of Estimation of Telephone Hold Time

Table 3. Customer satisfaction data for experiment one, showing the percentage score for the average of three customer satisfaction questions

<table>
<thead>
<tr>
<th>Choice</th>
<th>Short Wait Time</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Long Wait Time</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Row means</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0.92</td>
<td>0.05</td>
<td>0.90</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.89</td>
<td>0.04</td>
<td>0.88</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>0.68</td>
<td>0.04</td>
<td>0.82</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>240</td>
<td>0.37</td>
<td>0.05</td>
<td>0.51</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.71</td>
<td></td>
<td>0.77</td>
<td></td>
<td>0.74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No Choice</th>
<th>Short Wait Time</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Long Wait Time</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Row means</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0.82</td>
<td>0.04</td>
<td>0.89</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.79</td>
<td>0.05</td>
<td>0.68</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>0.53</td>
<td>0.05</td>
<td>0.61</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>240</td>
<td>0.40</td>
<td>0.04</td>
<td>0.45</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.64</td>
<td></td>
<td>0.66</td>
<td></td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Experiment one customer satisfaction data, by condition

the estimation error and the higher the customer satisfaction ($r_{(382)} = -0.22, p < .01$ and $r_{(382)} = 0.48, p < .01$). The correlations are both significant.

DISCUSSION OF EXPERIMENT 1

Table 4 summarizes the pattern of results from experiment 1. As Meyer, Shinar, Bitan and Leiser (1996) stated, “Factors that influence duration estimates may or may not influence preferences and vice versa” (p. 47), and the results of experiment 1 certainly bear out the assertion.

The purpose of the study was to test the effect of two variables on the subjective experience of time for participants waiting in a hold queue for a simulated consumer call center. The hypothesis concerning music choice (that wait time estimates
The Effect of Choice and Announcement Duration of the Estimation of Telephone Hold Time

Table 4. Pattern of results from experiment one. IV = independent variable, DV = dependent variable

<table>
<thead>
<tr>
<th>IV</th>
<th>DV</th>
<th>Estimate Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration of announcement</td>
<td>Customer Satisfaction</td>
<td>Marginally significant</td>
</tr>
<tr>
<td>Choice S</td>
<td>Estimate Error</td>
<td>Significant</td>
</tr>
</tbody>
</table>

The prediction that the ability to select the music type would result in a different subjective time experience than the no-choice condition was partially supported by the data. Participants indicated a higher level of satisfaction with the wait time when they were offered a choice of music types in the hold queue. However, contrary to the prediction, the effect of being able to choose a particular type of music on a participant’s estimates of the wait time was not significant.

The hypothesis concerning the duration of the hold queue announcement was also partially supported. It was predicted that a longer announcement would result in a shorter perceived wait time and higher customer satisfaction. The data show that participants significantly overestimate the wait time for the shorter announcement, but do not overestimate wait time for the longer announcement.

Table 4. Pattern of results from experiment one. IV = independent variable, DV = dependent variable
as well is some indication that the callers were perceiving, at some level and to some deleterious effect on the user experience, the longer duration announcements.

**EXPERIMENT 2: BETWEEN SUBJECTS TEST**

**Motivation for Experiment 2**

After experiment 1, there was some concern that the repeated-measures design, requiring as it did each participant to estimate times in 16 trials, may have led to artifactual effects that would not occur in the real world, where a customer support call is a relatively rare event. Zakay (1993) proposed that time could be assessed in two ways. In prospective mode, participants know in advance that they will be making a judgment of time for some event. In retrospective mode, participants do not know that they are going to make a time estimate until after the event occurs. Zakay postulated that prospective estimates would be subjectively longer because people devoted more attention to them.

Clearly, in experiment 1, after the first or first few tasks, the participants were in prospective time estimate mode. From this it can be inferred that, if Zakay (1993) was correct, time estimate errors would be less than they would be in the situation where the person does not go into the task expecting to estimate times. What is the real-world situation like? In a single, relatively rare call, are people typically NOT in prospective time estimate mode? Or could it be the case that, given how common it is to be put on hold, might many people be in prospective time estimate mode when dialing customer support, expecting to time the too-long wait for the impending call? In response to this concern, a between-subjects design was employed, where each participant had only one trial.

**METHODS: EXPERIMENT 2**

**Participants**

One-hundred twenty-eight individuals were paid (25 USD) to participate in a half-hour session. Age and gender distributions were similar to those of experiment one, as was the recruiting criteria.

**Apparatus**

The apparatus used was identical to that described in experiment 1.

**Procedure**

Experiment 2 was a partial replication of experiment 1, conducted as a between-subjects design. Each of 128 participants received one trial, with one of two levels of music choice (choice or no choice), and one of four levels of wait duration (30, 60, 120, and 240 sec); thus, there were eight conditions with 16 participants per condition. A trial was identical to that described in experiment 1.

**RESULTS: EXPERIMENT 2**

Table 5 summarizes the estimate error data (the difference between actual hold duration and participant’s estimates) for experiment 2. Across all four wait times, participants averaged 16.7 seconds overestimation, although they consistently underestimated the longest (240 seconds) condition. The trials on which participants were given no choice of music were overestimated by 0.8 seconds more, on the average.

A $2 \times 4$ (Choice $\times$ Hold Duration) between-subjects analysis was conducted for estimation error. There is a significant effect for hold duration only ($F_{(3, 120)} = 16.02, p < .0001$).
Table 5. Estimate Error in seconds for experiment two. Negative numbers indicate underestimation of length of time.

<table>
<thead>
<tr>
<th>Wait Time</th>
<th>Mean</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>8.9</td>
<td>2.9</td>
</tr>
<tr>
<td>60</td>
<td>3.4</td>
<td>11.6</td>
</tr>
<tr>
<td>120</td>
<td>57.8</td>
<td>4.2</td>
</tr>
<tr>
<td>240</td>
<td>-17.3</td>
<td>16.5</td>
</tr>
<tr>
<td>Overall</td>
<td>16.3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No Choice</th>
<th>Mean</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>8.1</td>
<td>7.1</td>
</tr>
<tr>
<td>60</td>
<td>5.7</td>
<td>15.6</td>
</tr>
<tr>
<td>120</td>
<td>46.3</td>
<td>3.8</td>
</tr>
<tr>
<td>240</td>
<td>-41.6</td>
<td>15.8</td>
</tr>
<tr>
<td>Overall</td>
<td>17.1</td>
<td></td>
</tr>
<tr>
<td>Grand Mean</td>
<td>16.7</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Customer satisfaction data for experiment two. Data are on a scale of 1 to 7, where 1 = Strongly Disagree, and 7 = Strongly Agree, that the participant was satisfied with the wait time.

<table>
<thead>
<tr>
<th>Wait Time</th>
<th>Mean</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>.69</td>
<td>.12</td>
</tr>
<tr>
<td>60</td>
<td>.94</td>
<td>.42</td>
</tr>
<tr>
<td>120</td>
<td>4.34</td>
<td>.34</td>
</tr>
<tr>
<td>240</td>
<td>2.47</td>
<td>.29</td>
</tr>
<tr>
<td>Overall</td>
<td>4</td>
<td>.61</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No Choice</th>
<th>Mean</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>.03</td>
<td>.23</td>
</tr>
<tr>
<td>60</td>
<td>.56</td>
<td>.41</td>
</tr>
<tr>
<td>120</td>
<td>3.31</td>
<td>.39</td>
</tr>
<tr>
<td>240</td>
<td>2.75</td>
<td>.35</td>
</tr>
<tr>
<td>Overall</td>
<td>4</td>
<td>.16</td>
</tr>
<tr>
<td>Grand Mean</td>
<td>4.39</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. Experiment two customer satisfaction data, showing the interaction between choice and ‘liked/didn’t like’
The customer satisfaction data are summarized in Table 6. There is a monotonic erosion of satisfaction across the four wait times. The choice condition yielded a higher average satisfaction score than the no-choice condition, 4.61 to 4.16.

A 2 × 4 (Choice × Hold Duration) between-subjects analysis was conducted for customer satisfaction. There is a significant main effect for hold duration \( F(3, 120) = 45.23, p < .0001 \) and no effect for choice \( F(1, 120) = 3.591, p = .061 \).

A 2 × 4 × 2 (Choice × Hold Duration × Like) between-subjects analysis was conducted for customer satisfaction (score from 1 to 7). The “Liked” variable is based on the questionnaire item that asked participants to rate the following statement: “I liked the music selection that I heard while I was on hold,” on a scale from 1 to 7. Only the effect of hold duration reached significance \( F(3, 112) = 26.56, p < .001 \). Plotting the interaction of Choice X Hold Duration (see Figure 4) shows an interesting result; those who got to choose their music were slightly more satisfied with the wait period when they did not like the music chosen, whereas those who got no choice were less satisfied with the wait period when they did not like the music. This could be interpreted as the propensity of people to take some of the blame for a negative experience on themselves when they get to choose.

**DISCUSSION OF EXPERIMENT 2**

The results found in experiment 2 were identical to those found in experiments 1. This means that experiment 2 failed to find a difference between prospective and retrospective assessments of time estimation in this variety of on-hold experiment that one would expect to see based on the work of Zakay (1993). The explanation assigned to this result is that because callers have significant experience with the delays associated with making calls to customer service, they are likely predisposed to be in a prospective mode when they make the call. That is, because they typically expect to be put on hold, their internal timers have already started when they make the call, in anticipation of the wait. For this task, it would seem that only naive callers would be able to use retrospective estimation in their first call. In a practical sense, however, it does not matter if retrospective estimation could be found in a subset of naive callers; the fact is that the data indicate that a majority of callers do have this expectation, and so designs (and the experiments to help determine those designs) must conform to the natural reality.

**GENERAL DISCUSSION**

Earlier studies (see the introduction for a comprehensive list) have clearly shown that the content of the wait time can influence the estimated duration. Some content is better than no content, at least when it comes to expressed customer satisfaction, but the effects on perceived length are more mixed. An Interactive Voice Portal is better than music or advertisements. The current work extends these findings along two important fronts. First, the length of the introductory announcement reduces apparent hold times, at least for the parameters specified in the current study. Second, providing a choice, even a minimal choice like the selection of music genre, increases caller satisfaction (but does not influence perceived wait time).

Memory load models would predict lower estimates for silence, and that has not been found in the vast majority of applied studies. The attention allocation model predicts that complex, interactive conditions would lead to shorter estimates. This was true for the Interactive Voice Portal (IVP), but not true for choice. A segment model predicts that more segments lead to longer estimates, but this was decidedly not so for the choice conditions, in which the opportunity to make a choice creates at least two additional segments. What do the results mean in the implementation
of real IVRs? Giving people a choice may not reduce the perceived duration of the wait, but people like it better anyway. Providing choice is a simple way to increase customer satisfaction in on-hold situations with a minimum of investment. Up to some yet-to-be-determined limit, using a longer recorded message does not count against the perceived wait time. Although there is likely an upper limit, the use of appropriately long announcement can reduce the time callers perceive they have been on hold.

Future research will need to identify the limits of announcement length as a means of reducing apparent wait time, as represented in Figure 5. Additionally, an attempt should be made to identify the effects of segments during the wait, rather than have that confounded with choice. For example, a choice/no choice study could be conducted where the no-choice condition has some additional segment (e.g., “You will hear some music. Tap the ‘M’ Key to continue.”) Finally, an attempt should be made to identify the additivity of the various factors. There were no significant interaction effects in the studies, suggesting that there should be a gain in additive effects by offering pleasant, choice-driven, interactive experiences, after long announcements.

CONCLUSION

While satisfaction with the wait period is just one component of the call center customer experience, any dissatisfaction due to on-hold delays can be mitigated by offering well-chosen content within that hold time. Offering users a choice of content and picking an optimal-length automated announcement can serve to reduce the apparent length of wait times and increase customer satisfaction. Finally, participants in these kinds of studies appear to be making time estimates using prospective memory for every trial, most likely due to previous experience with on-hold telephone situations.
ACKNOWLEDGMENTS

This article is dedicated to the memory of Dr. Robert R. Bushey; friend, colleague and co-author of this article.

REFERENCES


The Effect of Choice and Announcement Duration of the Estimation of Telephone Hold Time


APPENDIX A: EXPERIMENT ONE
TASK LIST

Task 1:
Call 9-241-6785 and ask “can I cancel my Caller ID service today?”

Task 2:
Call 9-241-6785 and ask “what is my Voice Mail password?”

Task 3:
Call 9-241-6785 and ask “how much does it cost to get DSL?”

Task 4:
Call 9-241-6785 and ask “when will the repair technician arrive at my home today?”

Task 5:
Call 9-241-6785 and ask “how much does it cost to get a second phone line?”

Task 6:
Call 9-241-6785 and ask “how much do I owe on my phone bill?”

Task 7:
Call 9-241-6785 and ask “how much does it cost to get Call Forwarding?”

Task 8:
Call 9-241-6785 and ask “can I have my phone service disconnected today?”

Task 9:
Call 9-241-6785 and ask “how much does it cost to get Caller ID?”

Task 10:
Call 9-241-6785 and ask “how much does it cost to get Call Notes?”

Task 11:
Call 9-241-6785 and ask “what address do I send my payment to?”

Task 12:
Call 9-241-6785 and ask “how much does it cost to get Call Blocking?”

Task 13:
Call 9-241-6785 and ask “is DSL service available in my neighborhood?”

Task 14:
Call 9-241-6785 and ask “can I get another copy of my bill sent to me?”

Task 15:
Call 9-241-6785 and ask “how do I block calls to my home phone?”

Task 16:
Call 9-241-6785 and ask “did you receive my phone bill payment?”

APPENDIX B: EXPERIMENT ONE SATISFACTION QUESTIONNAIRE

Estimate the amount of time you had to wait before speaking to customer service. Please be specific:

___________Minutes ___________Seconds

Please indicate how much you agree or disagree with each of the following statements by circling a number to the right of each statement.
<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Slightly Agree</th>
<th>Neutral</th>
<th>Slightly Disagree</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I liked the music selection that I heard while I was on hold.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>It seemed like the wait was too long before customer service came on the line to answer my question.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Listening to something different during the wait would have made the wait seem shorter.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I had to wait so long to get through that I wanted to hang up and try again at another time.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Overall, I was satisfied with my experience.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Chapter 7.15

Volkmar Lotz  
*SAP Research, France*

Luca Compagna  
*SAP Research, France*

Konrad Wrona  
*SAP Research, France*

**ABSTRACT**

The flexibility and dynamism of ubiquitous computing systems have a strong impact on the way their security can be achieved, reaching beyond traditional security paradigms like perimeter security and communication channel protection. Constant change of both the system and its environment demand adaptive security architectures, capable of reacting to events, evaluating threat exposure, and taking evolving protection needs into account. We introduce two examples of projects that contribute to meeting the challenges on adaptive security. The first focuses on an architecture that allows for adaptive security in mobile environments based on security services whose adaptation is guided by context information derived from sensor networks. The second addresses engineering aspects of secure ubiquitous computing systems through making security solutions accessible and deployable on demand and following emerging application-level requirements.

**INTRODUCTION**

A major challenge in securing ubiquitous computing systems is to cope with the increased flexibility and dynamism these systems show: the actual structure and behavior of a system at a particular point of time during its operation is not known in advance and depends on the physical and application context at that time. Consider
Towards Adaptive Security for Ubiquitous Computing Systems

a service-oriented architecture (cf., the chapter “Ubiquitous Services and Business Processes”), where a business service providing access to a resource—for example, performing a transfer to a bank account—is used in two applications being composed on demand and referring to a different business environment—for example, invoicing in a supply chain management application, and monthly employee payment in a human resources application. Otherwise, in a spontaneous interaction scenario (cf., the section “Ad-Hoc Interaction” of the chapter “Security for Ubiquitous Computing”) the number of participants in a communication as well as their roles change over time, with the particular type of interaction depending on the location of an entity, the networking capabilities available, and the particular features of the devices being used.

The flexibility of ubiquitous computing systems has a strong impact on security. The constant change of characteristics of both the system and its environment leads to different protection goals and exposure to threats over the system lifespan. Security thus needs to be adaptive: the security architecture, comprising both policies and mechanisms—as explicated in section “Sample Solutions” of the chapter “Security for Ubiquitous Computing”—needs to react to events, to evaluate the current threat exposure, and to take the actual protection needs into account. For example, the bank transfer service, being used in the human resources application environment, is likely to be subject to privacy regulations and confidentiality requirements protecting the information about an employee’s salary, whereas the same service, in a supply chain management environment, is probably required to provide strong traceability of transfers and to enforce the four-eye principle on their approval. The service might only be permitted to run on a mobile device in cases where the transferred amount does not exceed a given threshold, due to the increased vulnerability of wireless communication and the mobile environment to eavesdropping, unless it deploys strong authentication and encryption mechanisms.

The challenge of security solutions adapting to the system’s context occurs in development, deployment and operation of ubiquitous computing systems. In a service-oriented architecture, applications are designed on demand in a composite manner through orchestrating services, while taking advantage of the reuse of services in different application contexts. This asks for the application designer to specify the individual protection needs and security policies, and the service ecosystem (cf., the chapter “Ubiquitous Services and Business Processes”) and infrastructure to support the selection and the set-up of the appropriate security architecture as well as its configuration, all being accessible to an application designer not assumed to be a security expert. Such an effort needs to take into account that the system as well as its environment and protection needs might change over time; thus, events indicating a security relevant change—for example, change of a physical condition or location—need to be identified, the actions to be taken upon detection of the event—for example, the modification of the access control policy or the strength of encryption—need to be specified, and the monitoring of the events as well as the execution of the appropriate actions need to be enforced.

The remainder of this chapter introduces two examples of projects that contribute to meeting the challenges on adaptive security. The first focuses on an architecture that allows for adaptive security in mobile environments based on security services that—like application services in a service ecosystem—can be composed on demand to meet individual security requirements, and whose adaptation is guided by context information derived from sensor networks. The second addresses engineering aspects of secure ubiquitous computing systems through making security solutions accessible and applicable—in terms of integration and operation—on demand and following emerging application-level requirements.
MOBILE WORKERS’ SECURE BUSINESS APPLICATIONS IN UBIQUITOUS ENVIRONMENTS: THE MOSQUITO PROJECT

Background

The aim of a European research project—MOSQUITO (Mobile Workers’ Secure Business Applications in Ubiquitous Environment) (MOSQUITO, 2007; Wrona & Gomez, 2006)—was to provide a secure, trusted and ubiquitous access to business applications. The project developed an easy-to-use and flexible technical infrastructure, required so that mobile workers and their clients can perform daily business processes in a secure and collaborative manner. A particularly important segment of the teleworking market is mobile professionals, such as medical personnel, social workers, insurance agents, technicians, traveling executives, consultants and sales representatives. In today’s businesses, sensitivity to the needs of customers, closeness to their demands and tight relations with customers’ daily operational business processes is essential for success. This is why marketing, sales and production companies send out their mobile work forces in order to be as close to their customers as possible, if not permanently on their sites. However, this world is still characterized by obstacles such as inadequate communication facilities, incompatibilities between applications, and system boundaries that actually hinder interoperability between clients’ and customers’ systems more than they support it. This is equally true for security mechanisms that are insufficient, meeting neither the requirements of business users nor those of their clients—for example, when data is being used on systems that do not belong to the owner of the data.

MOSQUITO Architecture

Figure 1 gives an overview of the MOSQUITO architecture. The platform provides basic security functionalities like protecting key material, authenticating a user connected to a device, or ensuring the integrity of the sensitive data. The platform also gives access to local context information through sensors. The heart of the MOSQUITO architecture is a middleware including a set of security services. Finally, applications are either written on top of the middleware or, in case of legacy applications, are adapted to use this middleware. Application-specific policies are defined to configure the middleware.

At a high level, the developer of a MOSQUITO-based system may be confronted with the assembly of a workflow together with some basic initialization of the MOSQUITO components.

Figure 2 depicts the interactions taking place between the various components that implement this workflow behind the scenes and how they have to be set up to properly interoperate with a service-based application. The interface to the MOSQUITO framework from the application developer’s point of view consists of the message filters, Context Information Acquisition and Trust Service (CATS) and the discovery service.

Figure 1. The MOSQUITO architecture
MOSQUITO Scenario

In the following, we describe a typical scenario for how the MOSQUITO framework is used in the healthcare application domain, in particular, elderly care. A patient Bob, who suffers from restricted mobility and stays more or less at home, has subscribed to a monitoring service that ensures assistance in case of illness. Bob’s health status is monitored by several sensors, for example, temperature or pulse sensors. A monitoring application analyzes the data and in case of irregularities (high temperature or high pulse rate) a workflow is initiated to schedule a house visit to the patient and provide medication if necessary.

The first step of the triggered workflow is to select physicians who are able to make a house visit. The selection is based on the patient’s preferences and needs such as language and location. The first physician gets a notification via mobile phone about the house visit to make, which he has to accept. If he refuses the next doctor is notified. Once a doctor has accepted the notification, he is responsible for scheduling an appointment with the patient and for making the house call on time.

The physician visits the patient and conducts an examination of the patient. The physician may require access to the patient’s medical data. The patient’s data are stored on a Medical Information Portal (MIP). The physician uses a Web application on his laptop or his mobile device (PDA) to view the medical data. Access to the medical data is controlled on a context-aware way. In particular, the physician gets access to the patient’s data only when he has accepted the house call notification (workflow is in the examination state) and when he is close to the patient. Proximity is evaluated based on the physical distance (GPS-based) between the patient and the physician. Both patient and doctor are equipped with a mobile phone that can provide the GPS location.

The examination may end with the prescription of a drug. The prescription is issued electronically via Web application and is sent automatically to the nearest pharmacy. A pharmacist will take care
of the medication and prepare or order the drugs if necessary. As soon as the medication is ready for dispatch, the pharmacist uses a Web application to confirm and the workflow continues.

In the next step, a social worker is selected (based on availability) and gets a notification via mobile phone. The notification includes information about where to pick up the drugs and their delivery address. The social worker has to explicitly accept the notification and, by doing so, he commits to carrying out the required tasks.

A special case is the reassignment of the doctor who has already agreed to perform the house call. In our scenario we assume that the doctor gets an emergency call and is therefore unable to make the house visit in time. During this scenario the issued workflow task is reassigned to the next available doctor.

**Pervasive Workflow**

The demonstrator that has been developed throughout the MOSQUITO project outlines the collaboration between actors from the healthcare application domain. As part of the MOSQUITO framework, we also developed a workflow management system—so-called pervasive workflow (Montagut & Molva, 2006a; 2006b)—easing the dynamic collaboration of partners across organizational boundaries. The pervasive workflow engine deployed on each site involved in a distributed collaboration is in charge of orchestrating the overall process execution. Compared to usual agent-based solutions, this model lowers the risk of malicious code hidden in messages exchanged during a workflow instance. As opposed to centralized workflow management systems, the distributed execution of workflows raises security constraints due to the lack of dedicated infrastructure assuring the management and control tasks of workflow instances. As a result, basic security features such as compliance of the workflow execution with the predefined plan are no longer assured. We categorize the security requirements we identified for distributed workflow systems into three main categories: authorization, proof of execution and workflow data protection. In order to achieve the enforcement of these security requirements, we capitalized on security mechanisms ranging from onion encryption in the specification of execution proofs to ID-based cryptography for the workflow policy enforcement. At workflow design phase, a workflow policy is defined to specify the credentials required by prospective business partners, who could be assigned to the workflow tasks. The transactional model we designed for the pervasive workflow model was implemented using Web services (WS) framework. On one hand an extension of an OWL-S matchmaker integrating the theoretical results was proposed. On the other hand, we implemented a proof of concept of the suggested transactional protocol capitalizing on the WS-Coordination specification. Use of contextual information during the workflow execution was demonstrated by including doctor’s availability as one of the parameters.

**Evaluation**

Basic security mechanisms, for example, integrity, confidentiality, and availability, are common grounds for any business scenario. However, healthcare applications introduce specific security requirements such as non-repudiation or accountability. Our work towards the integration of security mechanisms to match these requirements included two aspects. First, our demonstrator makes use of the special security component (CryptoTerminal) in order to assign business partners to workflow tasks and in order to enable signing of relevant workflow documents. This is, for instance, the case for electronic prescriptions that are exchanged throughout the healthcare workflow. This enables, on the one hand, to gather commitments from business partners during the partner discovery procedure as soon as they accept to be involved in a given workflow instance, and, on the other hand, to identify the issuer of
Towards Adaptive Security for Ubiquitous Computing Systems

critical documents, as those are digitally signed. Second, we have designed security mechanisms to ensure both runtime and “a posteriori” verification of compliance with the actual plan of execution by means of a policy-based approach. The solutions that we currently provide are basic mechanisms that partially answer the required security mechanisms, which are accountability and non-repudiation. Designing and implementing mobile and secure logs would have been a perfect means to solve the limitations of our current solutions yet this research effort was outside of the MOSQUITO project main focus.

Context-awareness is a key feature in MOSQUITO that comes in several flavors. Context may be used directly by applications that want to adapt to their environment, in which case its accuracy has an impact on the correct operation of an application. However, context may also be used by security mechanisms, in particular at the access control level, in which case accuracy is fundamental to prevent security breaches that may threaten the availability of services or the confidentiality or integrity of their data. This in particular requires the context acquisition process to be able to check how trustworthy the acquired context information is. The integration of context information into MOSQUITO framework relies on two components: the Context Information Acquisition and Trust Service (CATS), and the Context Information Service (CIS). The CATS is used in order to figure out where to acquire the CI among a set of available CIS servers. The CATS may either select an appropriate CIS out of a static list of known and reliable services, or rely on the Secure Discovery Service to locate CI Services based on a dynamic and semantically rich profile. The CATS also provides basic algorithms for evaluating the trustworthiness and reliability of the context information acquired before providing it to interested applications. The prototype implementation of the CATS and the CIS was equipped with two trust evaluation methods: a PKI-based validation of context information, and a voting-based validation of context information, based on the comparison of context acquisition from a set of redundant CI Services.

The MOSQUITO framework supports a wide range of underlying communication technologies and execution platforms, including wireless networks and mobile devices. Nevertheless, implementing a service-oriented architecture, and heavily relying on Web services has twofold consequences: on one hand, it ensures an easy integration with existing service ecosystems and enterprise systems and makes the framework easier to use for developers familiar with standard Web services technology. On the other hand, it puts higher requirements regarding the communications bandwidth and computing power of end-user devices. In particular, it is currently not possible to execute the full MOSQUITO framework on a small mobile device. In our demonstration and experiments we have relied on laptops and tablet PCs, limiting the use of the phone to personal authentication and signature. However, high-end mobile phones and PDAs already support Web services, and we believe that in the near future, the execution of a complete MOSQUITO framework on much smaller devices should be possible. The bandwidth requirements of the MOSQUITO applications are compatible with most of the existing technologies, including GPRS, 3G, and WLAN. The delay induced by the processing and communication is noticeable in our demonstrator, but we believe that it could be significantly reduced by optimizing the implementation and deployment configuration.

The MOSQUITO framework provides a robust and innovative middleware concept for context-aware business applications in mobile environments. Although it does support the dynamic formation and execution of a workflow, the implemented interaction scheme is however not fully spontaneous—that is, although it does permit dynamic utilization of WLAN or mobile
Towards Adaptive Security for Ubiquitous Computing Systems

networks, currently it does not support fully ad hoc communications. In every ubiquitous environment, a significant amount of implementation work is currently related to the development of an interface between middleware (i.e., CIS) and the physical sensor or sensor network. An easy and seamless integration of sensors and sensor networks with services is still an open issue.

Conclusions

The MOSQUITO project investigated into a middleware supporting secure, collaborative, and ubiquitous access to applications for mobile workers. Key features of MOSQUITO include the support of context-awareness through sensor networks and pervasive workflows that allow distributing workflows across a set of devices without central control. MOSQUITO has been validated by means of a prototype supporting an elderly care scenario, where context information is used to control access to medical data through a mobile device, and a pervasive workflow facilitates the coordination of the actions of the different stakeholders being involved.

SYSTEMS ENGINEERING FOR SECURITY AND DEPENDABILITY: THE SERENITY PROJECT

Background

There is a common understanding that security solutions built into systems after the fact (of design decisions made and system development done) turn out to be ineffective or even inappropriate. Security needs to be designed into a system from the very beginning, leading to a discipline of security engineering supporting all phases of secure system construction from security requirements analysis to security mechanisms design, and stretching out to system deployment and operation. While there is a rich body of design principles, building blocks, and case studies available, for example, Anderson (2001), their integration in development and deployment environments is still immature. The systems engineer is, thus, required to be security aware, and the assessment of the security contribution of a mechanism and its appropriateness for the requirements at hand are left to the security experts.

The engineering challenge is amplified when considering ubiquitous computing: flexibility and dynamism demand adaptive security architectures, as motivated in the introduction, and new security solutions have to be deployed on the fly. Thus, a designer or developer needs an immediate link between emerging security requirements and implemented solutions to satisfy them. The highly distributed nature of and decentralized control in ubiquitous computing systems demand the consideration of interoperability and the heterogeneity of infrastructures. For instance, in a service ecosystem, a security mechanism protecting a service or a resource needs to interact with other security mechanisms owned by different entities, taking interoperability, consistency of security properties, and potential side effects into account. Still, dynamic behavior does not allow evidence to be provided that a security architecture effective at one point of time will still be appropriate at some future point of time, thus requiring continuous monitoring of the security status of the running system.

The Serenity Approach

Following the above, ubiquitous computing demands the dynamic application of security expertise (Mana et al., 2007). The Serenity project (Serenity, 2006) provides a framework to support this goal through:
Towards Adaptive Security for Ubiquitous Computing Systems

- Systematic capture of security expertise in terms of enhanced security and dependability (S&D) patterns and integration schemes, linked to ready-to-integrate implementations of security solutions, for example, in terms of security services.
- Access to expertise through a S&D library integrated into the development and run-time environment.
- A monitoring framework, able to capture context information and triggered by events being specified in the S&D patterns.
- Deployment support for S&D solutions based on S&D patterns and integration schemes, taking information about the system architecture and its context into account.
- An end-user interface, allowing the specification of S&D requirements and their mapping to solutions and implementations through the S&D library, and a system interface allowing retrieval of context and interaction across Serenity-enabled systems.

Figure 3 shows the framework architecture resulting from this approach (Sánchez-Cid et al., 2006). The Serenity framework has to be instantiated each time it has to be applied to a concrete scenario, for example, an application, a device, or a security domain. After specifying the security requirements by the user (i.e., a designer, a developer, an administrator, or an application user), the framework, based on the context information retrieved, suggests appropriate security solutions through S&D patterns and integration schemes, allowing reasoning about alternate options, constraints, and interdependencies. This investigation leads to a final selection of a set of security solutions comprising the system’s security architecture and its deployment. Note that the framework is aware of the deployed architecture through a model describing the actual security configuration, allowing exploitation of this information upon monitoring the system at run-time or even exchanging it with other framework instances.

Patterns and Integration Schemes

The notions of S&D patterns and integration schemes are crucial to the approach, since they contain the information about an S&D solution’s effects, the link to satisfied requirements and implementations, as well as the specification of potential interferences when combining different solutions.

Patterns were first introduced in software engineering (Beck & Cunningham, 1987; Buschmann et al., 1996). This approach was then adopted to schematically describe security solutions in an informal way, for example, in Fernandez (2000; 2004), Schumacher (2003), and Schumacher et al. (2006). Security patterns describe a security solution in terms of the problem context (an abstract description of assumptions made and not to be confused with the use of the notion of “context” above), a problem statement, a solution description and implementation advice. They explicitly
Towards Adaptive Security for Ubiquitous Computing Systems

Serenity significantly extends the notion of patterns in several dimensions:

- Serenity patterns refer to properties that are provided upon deployment of the pattern (assuming all specified constraints are met, or enforced by the framework). This allows us to link user requirements to patterns, since they are expressed in terms of properties as well.
- Serenity patterns, in general, refer to an implementation, that is, program code, that is ready to be integrated into a given system implementation through a defined interface.
- Serenity patterns include obligations on the run-time of the system in terms of events to be monitored and controls to be imposed. The obligations are enforced through the framework.
- Serenity patterns are parameterized and can be hierarchical, thus increasing their flexibility and applicability.
- Security solutions in Serenity patterns are verified, preferably through formal methods. This asks for much more precise specification of a solution than the standard pattern approach does.
- A particular type of Serenity patterns, called integration schemes, allow capturing potential interference in case of patterns being combined, as well as imposing constraints with respect to their deployment.

Figure 4 shows the modeling artifacts used in Serenity. In addition to patterns, properties and implementations, the figure introduces classes as an additional abstraction level. This allows consideration of the fact that security requirements are often given in terms of general notions (like confidentiality, integrity, non-repudiation),

Figure 4. Serenity modeling artifacts: Classes, patterns, and implementation
Towards Adaptive Security for Ubiquitous Computing Systems

with these notions then being refined according to individual interpretations.

Evaluation

The Serenity idea is ambitious and likely to change the way secure systems are developed, deployed and operated. The Serenity library of patterns and integration schemes facilitates a systematic approach to accessing security expertise, based on specifications of security properties covering different layers of abstraction, from business level to network and devices level. The Serenity framework provides means to monitor and reason about a deployed security architecture and to adapt it if necessary. Thus, major requirements on security in ubiquitous computing systems are met: accessibility of security solutions to the non-expert developer, and automated support for adapting security solutions to system evolution.

The key success factors for the Serenity approach include the coverage and quality of the library entries, the integration of the specification means across the abstraction levels, the control of security architectures through the framework, and the interoperability of the security solutions. The latter is particularly challenging in heterogeneous infrastructures that are characteristic to ubiquitous computing. To ensure the feasibility of the approach, Serenity is evaluated through a number of case studies, including such varying domains as e-business, sensor networks, mobile communications, e-government, and air traffic management.

Conclusions

Serenity addresses the problem of security engineering for ubiquitous computing systems through capturing security expertise by means of enhanced patterns and integration schemes as well as providing a framework for components and devices supporting the systematic development, deployment and control of security solutions. Since the framework is aware of the deployed security architecture, feedback from security monitoring can be used to adapt the architecture according to new requirements or system evolution. Serenity’s feasibility critically depends on meeting the challenge of interoperability of the solutions made available through the patterns of the Serenity library.

REFERENCES


Montagut, F., & Molva, R. (2006a). Augmenting Web services composition with transactional requirements. In Proceedings of the IEEE Inter-
Towards Adaptive Security for Ubiquitous Computing Systems

national Conference on Web Services. Chicago, USA (pp. 91-98).


This work was previously published in the Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises, edited by M. Mühlhäuser and I. Gurevych, pp. 574-584, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 7.16

Trusting Computers Through Trusting Humans:
Software Verification in a Safety-Critical Information System

Alison Adam
University of Salford, UK

Paul Spedding
University of Salford, UK

ABSTRACT

This article considers the question of how we may trust automatically generated program code. The code walkthroughs and inspections of software engineering mimic the ways that mathematicians go about assuring themselves that a mathematical proof is true. Mathematicians have difficulty accepting a computer generated proof because they cannot go through the social processes of trusting its construction. Similarly, those involved in accepting a proof of a computer system or computer generated code cannot go through their traditional processes of trust. The process of software verification is bound up in software quality assurance procedures, which are themselves subject to commercial pressures. Quality standards, including military standards, have procedures for human trust designed into them. An action research case study of an avionics system within a military aircraft company illustrates these points, where the software quality assurance (SQA) procedures were incommensurable with the use of automatically generated code.

INTRODUCTION

They have computers, and they may have other weapons of mass destruction. Janet Reno, former US Attorney General

In this article our aim is to develop a theoretical framework with which to analyse a case study where one of the authors was involved, acting as an action researcher in the quality as-
This involved the production of software for aeroplane flight systems. An interesting tension arose between the automatically generated code of the software system (i.e., ‘auto-code’—produced automatically by a computer, using CASE [Computer Aided Software Engineering] tools from a high level design) and the requirement of the quality assurance process which had built into it the requirement for human understanding and trust of the code produced.

The developers of the system in the case study designed it around auto-code—computer generated software, free from ‘human’ error, although not proved correct in the mathematical sense, and cheaper and quicker to produce than traditional program code. They looked to means of verifying the correctness of their system through standard software quality assurance (SQA) procedures. However, ultimately, they were unable to bring themselves to reconcile their verification procedures with automatically generated code. Some of the reason for this was that trust in human verification was built into (or inscribed into [Akrich, 1992]) the standards and quality assurance procedures which they were obliged to follow in building the system. Despite their formally couched descriptions, the standards and verification procedures were completely reliant on human verification at every step. However these ‘human trust’ procedures were incompatible with the automated production of software in ways we show below. The end result was not failure in the traditional sense but a failure to resolve incommensurable procedures; one set relying on human trust, one set on computer trust.

Our research question is therefore: How may we understand what happens when software designers are asked to trust the design of a system, based on automatically generated program code, when the SQA procedures and military standards to which they must adhere demand walkthroughs and code inspections which are impossible to achieve with auto-code?

The theoretical framework we use to form our analysis of the case study is drawn from the links we make between the social nature of mathematical proof, the need to achieve trust in system verification, the ways in which we achieve trust in the online world, the methods of software engineering, and within that, the software quality movement and the related highly influential domain of military standards.

In the following section we briefly outline the social nature of mathematical proof. The next section discusses the debate over system verification which encapsulates many of the ideas of mathematical proof and how such proofs can be trusted by other mathematicians. The article proceeds to consider ‘computer mediated’ trust, briefly detailing how trust has been reified and represented in computer systems to date, mainly in relation to the commercial interests of e-commerce and information security. Trust is particularly pertinent in the world of safety-critical systems, where failure is not just inconvenient and financially damaging, although commercial pressures are still evident here, but where lives can be lost. The model of trust criticised by e-commerce critics is more similar to the type of trust we describe in relation to safety-critical systems, than one might, at first, expect. Understandably, we would like to put faith in a system which has been mathematically proved to be correct. However computer generated proofs, proofs about correctness of computer software, and automatically generated code are not necessarily understandable or amenable to inspection by people, even by experts. The question then arises of whether we can bring ourselves to trust computer generated proofs or code, when even a competent mathematician, logician, or expert programmer cannot readily understand them.

Following this, we describe the evolution of software development standards and the SQA movement. We argue that the development of quality assurance discourse involves processes of designing human ways of trusting mathematical evidence into standardisation and SQA. Military
standards are an important part of the SQA story, having consequences far beyond the military arena. Standards are political devices with particular views of work processes inscribed (Akrich, 1992) in their design. We note the way that military standards, historically, moved towards formal verification procedures only to move back to rely more on ‘human’ forms of verification such as code walkthroughs and inspections in the later 1990s. The story is shot through with a tension between finding ways to trust the production of information systems and finding ways to control them. Formal methods, based on mathematical proof offer the promise of control, but only if we can bring ourselves to trust a proof generated by a machine rather than a proof constructed by another person. We present the background to the case study in terms of a description of the complex ‘post cold war’ military and commercial environment. This is followed by a description of the action research methodology employed in the project, an outline of the case study and an analysis of the case study findings in terms of our theoretical framework. In the conclusion we briefly note that mathematicians and others are gradually finding ways of trusting computers.

THE SOCIAL NATURE OF MATHEMATICAL PROOF

At first sight, the concept of mathematical proof appears to be relatively simple. The idea of a logical and rigorous series of steps, leading from one or more starting positions (previous theorems or axioms) to the final conclusion of the theorem seems to be the basis of mathematics. The concept of mathematical proof leading inexorably to true and incontrovertible truths about the world is very compelling. It is not surprising that we would like to apply the apparent certainty and exactness of mathematical approaches to computer programming. However if we consider briefly how agreement on mathematical proof and scientific truth is achieved by communities of mathematicians, then the social and cultural dimension of proof, as an agreement amongst trusted expert witnesses, reveals itself.

With the epistemological and professional success of mathematical proof, many of the cultural processes which go into making a proof true sink from consciousness and are only rendered visible in times of dispute; for example as in claims to the proof of Kepler’s conjecture or Fermat’s last theorem (Davies, 2006; Kuhn, 1962; Singh, 1997). Only on the margins then do we call into question our ability to trust these people when a mathematical proof cannot be agreed to be true by an expert community of mathematicians, as sometimes happens.

The apparently pure and abstract nature of mathematical proof fairly quickly breaks down when we inspect it more closely. In particular, when there is disagreement about a proof, the nature of proof is revealed as a social and cultural phenomenon; the matter of persuading and convincing colleagues. DeMillo, Lipton, and Perlis (1977, p. 208) wrote

Mathematicians talk to each other. They give symposium and colloquium talks which attempt to convince doubting (sometimes hostile) audiences of their arguments, they burst into each others’ offices with news of insights for current research, and they scribble on napkins in university cafeterias and expensive restaurants. All for the sake of convincing other mathematicians. The key is that other mathematicians are inclined to listen!

This traditional approach towards mathematical proof, which could be described as one of persuasive rigorous argument between mathematicians leading to trust, is not the only way to address the idea of proof. A quite different approach appeared in the 1950s and was based on the work on logic developed by Bertrand Russell and others in the 1930s and used the newly invented electronic computer. This new
logic-based approach was not dependent on the computer, but the computer’s speed and accuracy had a major impact on its application to the proof of theorems in replacing the persuasive rational argument of competent mathematicians with a formal approach which sees any mathematical proof as a number of steps from initial axioms (using predicate logic), to the final proof statement (based purely on logical inference) without the requirement of a human being.

Many proofs can be completed by either method. For instance, many persuasive rigorous argument proofs can be converted to formal proofs (MacKenzie, 2004). It should be emphasised, however, that there is a real difference between the two types of proof. We are not simply talking about a machine taking on the role of a competent mathematician. Some proofs which are readily accepted by mathematicians rely on arguments of symmetry and equivalence, analogies, and leaps of imagination, which humans are very good at understanding but which a formal logic approach cannot replicate. Symmetry and analogy arguments of this type cannot be established by formal methods based on logical progression because symmetry relies on understanding semantics and cannot be gleaned from the syntax of a proof.

Whereas the persuasive rigorous argument, the ‘human’ approach, has been used for thousands of years, the formal or ‘computer generated’ approach has been in use for only about half a century. Clearly, the two methods are not treated in the same way by the expert community of mathematicians. With a rigorous argument type of proof, although one may expend much energy convincing one’s colleagues of the validity of the proof, the potential for coming to agreement or trust of the proof is there. Essentially, in trusting that a mathematical proof is correct, mathematicians are demonstrating their trust in other competent mathematicians. However, expert mathematicians clearly have trouble bringing themselves to trust computer proofs, for good reason, as a computer cannot explain the steps in its reasoning (Chang, 2004).

COMPUTER SYSTEM VERIFICATION: TRUST AND THE SOCIAL

The preceding section contrasted the use of computer technology in a claimed proof: the formal method and the human ‘rigorous argument’ approach to proof. Although this is not the same thing as the proof or verification of a computer system itself, in other words the formal, computer generated proof that the computer system matches the specification, the question of whether we can trust the computer is exactly the same.

The idea of proof or verification of a program is quite different from simply testing the program. Typically, a large suite of programs might have thousands or millions of possible inputs, and so could be in many millions or even billions of states. Exhaustive testing cannot be possible. If a computer system is to be used in the well-funded and high-profile military field to control a space craft, aeroplane, or a nuclear power station, it is highly desirable if the system can be actually proved to be correct, secure, and reliable. Since testing, although vital, can never prove the system’s correctness, more mathematical methods involving the notion of proof became of great interest in the late 1960s and have remained so ever since.

In fact the history of the verification of computer systems echoes that of mathematical proof, with basically the same two approaches: those who support the rigour of formal methods and those who believe that the purely formal, mechanised proof lacks the crucial element of human understanding (Tierney, 1993). In a paper to an ACM Symposium, DeMillo et al. (1977) argued that the two types of proof were completely different in nature, and that only the persuasive rigorous argument proof with its strong social aspect will
 ultimately be believable and capable of earning trust

COMPUTER-MEDIATED TRUST

In ethical terms, trust is a complex phenomenon and is essentially a human relationship (Nissenbaum, 1999; Stahl, 2006). We think of trust in terms of a trustor who does the trusting and a trustee who is trusted. The trustee does not of course have to be human, but Nissenbaum (1999) suggests that the trustee should be a being to whom we ascribe human qualities such as intentions and reasons, what might be termed an ‘agent.’ Trust allows meaningful relationships and a vast range of intuitions to work. Nissenbaum (1999) argues that when we are guaranteed safety trust is not needed: ‘What we have is certainty, security, safety—not trust. The evidence, the signs, the cues and clues that ground the formation of trust must always fall short of certainty; trust is an attitude without guarantees, without a complete warrant.’

Intrusive regulation and surveillance are attempts at control and bad for building trust.

This generalised definition of trust clearly maps onto our description of mathematicians trusting proofs. They may not have complete certainty over the correctness of a mathematical proof, but they have good reason to trust a competent member of the community of expert mathematicians. Therefore they can trust the proof supplied by such a person.

Understandably, there has been much interest in trust in the online world, both in terms of online security and trust in e-commerce transactions. Nissenbaum (1999) suggests that excessive safety controls, say in e-commerce, may encourage participation but they limit experience: ‘Through security we may create a safer world, inhospitable to trust not because there is distrust, but because trust cannot be nourished in environments where risk and vulnerability are, for practical purposes, eradicated.’

Stahl’s (2006) take on trust in e-commerce shows another example of the intangible human nature of trust, which has become reified and commodified, so that it can be measured and exchanged in machine transactions. Like Nissenbaum (1999), Stahl points to the way that a trustor does not have complete control over a trustee; vulnerability and uncertainty must be accepted in a trusting relationship. This of course includes business transactions, and is especially important in e-commerce as many of the traditional ways of developing trust are absent from online transactions. Trust becomes a way of generating profit; small wonder that trust, including technological ways of creating trust and maintaining it, has been of so much interest in e-commerce. In the world of e-commerce research, trusts lose its relational aspects and becomes a form of social control. ‘If trust is limited to calculations of utility maximisation in commercial exchange, then most of the moral underpinnings of the mechanisms of trust become redundant. Trust changes its nature and loses the binding moral quality that it has in face-to-face interaction.’ (Stahl, 2006, p. 31)

Although, on the face of it, Nissenbaum’s and Stahl’s arguments on the problems of online trust in e-commerce are not the same as the issue of trust described in the body of this article, there are important congruencies which are very directly applicable to our characterisation of trust. Whether it is a human trusting another human or an expert mathematician trusting another expert mathematician to supply an accurate proof, the same relationship between trustor and trustee obtains.

For Nissenbaum and Stahl, the issue is what happens to trust when it is commodified within an online relationship. In other words, what happens when the human-trusting-human relationship is mediated by technology? In this article we also consider what happens when the human-trusting-human relationship—in terms of a human trusting another human’s mathematical proof, or computer program—is replaced by a human
having to trust a machine. Of course, in this truster-trustee relationship, the trustee, that is, the machine, cannot be understood in the way that another person can be.

The pressure to create computer-mediated trust is completely bound up with commercial pressures. The maximisation of profit drives the reification of trust in e-commerce. Similarly in the world of military avionics we describe, it is the commercial pressure of building systems more cheaply and faster which provides the impetus to turn over proofs, testing of programs, and automatic generation of code to a machine.

A third aspect of similarity between Stahl’s and Nissenbaum’s view of computer-mediated trust and ours relates to the tension between trust and control. This is clearly present in the debate over trust in e-commerce. But it is also present in software quality discourse as we discuss below.

In the following section we briefly discuss some of the ways in which human trust has traditionally been built into procedures designed to verify program correctness, and how this can be seen to mirror an ideal group of mathematicians agreeing upon a mathematical proof.

**BUILDING TRUST INTO A COMPUTER SYSTEM**

We argue that, historically, much of the development of the software engineering discipline can be understood in terms of the development of procedures, through which we can convince ourselves to trust, and control, the development of information systems and the production of software. For instance, Myers’ (1979) classic book on software testing explores the topic of human testing in detail, justifying methods such as formal code inspections and code walkthroughs. The differences between the two methods depend on different usages of the terms ‘inspection’ and ‘walkthrough,’ but the important point is that both involve a small group of professionals carefully reading through code together. We argue that this can be viewed as an imitation of the social (persuasive rigorous argument) form of proof described earlier where ‘mathematicians talk to each other’ in symposia and colloquia and so on (DeMillo et al., 1977). The original programmer should be in the group, analogous to the mathematician demonstrating a proof or principle to expert colleagues. The aim (as originally suggested by Weinberg [1971]—an ‘egoless’ approach) is to discover as many errors as possible rather than to try to demonstrate that there are none. So the team is to act as an idealised group of ‘Popperian’ scientists looking for ‘refutations’ (Popper, 1963). Under such an approach, one can never be entirely sure that the code is correct. But, as the walkthrough proceeds, the original programmer and the code inspection team can gradually come to trust the code as bugs are weeded out and fixed.

Myers claims positive advantages of code inspections and walkthroughs, including the value of the original programmer talking through the design (and thus spotting the errors). He also notes the ability of human testers to see the causes and likely importance of errors (where a machine might simply identify symptoms) and also the likelihood that a batch of errors will be identified simultaneously. Also the team is able to empathise with and understand the thought processes of the original programmer in a way which a machine arguably cannot. Importantly, the team can be creative in its approach. In working together they also, inevitably, form something of a sharing and trusting community (even if it is disbanded after a day or two).

The lesson gleaned from human verification techniques, such as walkthroughs and code inspections, is that these have been regarded, for some time, as reliable, if not exhaustive, ways of ensuring reliability of software.
SOFTWARE QUALITY ASSURANCE AND MILITARY STANDARDS FOR SOFTWARE

The software verification techniques of code walkthroughs and inspections are important parts of the armoury SQA. Effectively, we argue that SQA is a branch of software engineering which formalises and standardises the very human methods of trust, and ultimately control outlined above, which we need to build into software engineering procedures. The SQA movement is an important part of the story of the growth of software engineering because of its quest for rigour and control of potentially unruly programs and programmers.

First of all, SQA offers a promise of rational control over software, the software development process, and those who produce software. Software quality criteria include features for directing, controlling, and importantly, measuring the quality of software (Gillies, 1997). ‘Qualification’ is achieved when a piece of software can be demonstrated to meet the criteria specified in these quality procedures. An important aspect of SQA involves demonstrating that software meets certain defined independent standards.

The development and adherence to software standards is a very important part of the story of SQA. Generic industry standards are available, but also of much interest—particularly for the case study set out later in the article—are military standards. Indeed, the defence industry is so influential that Tierney (1993) argues that military standards influence software engineering far beyond applications in defence. Hence military standards are a very important part of SQA, and ultimately are important in formalising ways in which designers of computer systems can come to trust the systems and the production of correct software.

A number of military standards have been developed to regulate and control the use of software in defence applications. For instance, US standards DOD-STD-2167A (1988), MIL-STD-498 (1994), and ISO/IEC 12207 (1995) respectively established the requirements for software development and documentation in all equipment to be used by the US military (and effectively that of all Western armed forces), introduced object oriented development (OOD) and rapid application development (RAD), then broadened the scope of international standards to include acquisition and maintenance. (DSDM Consortium, 2006).

The relevant UK standard 00-55, (MoD, 1997) Requirements for Safety Related Software in Defence Equipment, was published in 1997 and echoes much of MIL-STD-498, but moves the discussion on provably correct software in a particular direction. At first sight, this seems highly significant to the current argument, because it clearly expressed a preference for formal methods, in other words mathematical procedures whereby the software is proved to be correct by a machine (MacKenzie, 2001).

Tierney (1993) argues that the release of UK Defence Standard 00-55 in draft in 1989 had the effect of intensifying the debate over formal methods in the UK software engineering community. It devoted as much space to regulating and managing software development labour processes as the techniques and practices to be used for formal designs. This reinforces our argument that SQA is concerned with control of work processes and those who perform them, the software developers. On the one hand, many argued that mathematical techniques for software development and verification could only ever be used sparingly, as there simply was not enough suitable mathematical expertise in most organisations and it increased software quality at the expense of programmer productivity. On the other side, those from a more mathematical camp argued that there was commercial advantage in proving software correctness as errors could be
trapped earlier in the software development cycle (Tierney, 1993, p. 116).

Designed into the MoD (UK Ministry of Defence) standard was a view of safety-critical software as an important area of regulation and control. Some of the reason for this was a change in its own organisation from the 1980s. The UK government sought to open up work traditionally done in-house by the MoD in its own research establishments to private contractors (Tierney, 1993, p. 118). Given that it had to offer its software development to the private sector, it built in ways of controlling it within its defence standards (Tierney, 1993, p. 118). Further political impetus was offered by the introduction of consumer protection legislation in the UK in the late 1980s which required software developers to demonstrate that their software had not contributed, in the event of an accident enquiry, and that they had demonstrably attended to safety. Thus we can see that in Def Stan 00-55, politics, in the shape of the MoD’s need to open up software development to the private sector and also to avoid being held responsible for inadequate software in the event of an accident, played an important role.

However, more significantly, this document has itself been superseded in 2004 by (draft) standard 00-56 (MoD, 2004). Def Stan 00-55 has now become obsolete. The changes involved in Def Stan 00-56 are of great interest, in that the preference for formal method is lessened. In the new standard, it is accepted that provably correct software is not possible in most cases and that we are inevitably involved in a human operation when we attempt to show that code is reliable in a safety-critical environment. Without a more detailed consideration of the history of formal methods in the UK over the last decade, which is beyond the scope of the present article, a strong claim that the move back to more human methods of verification might be difficult to sustain. Nevertheless it is interesting to note the way that Def Stan 00-5, with its emphasis on formal approaches and attendant onerous work practices, has been consigned to the history books with a clear move back to human verification.

**CASE STUDY CONTEXT**

The case study relates to a large European military aircraft company (MAC) with which one of the authors was engaged as a researcher in a joint research project, lasting around three years, during the mid to late 1990s. A high proportion of the senior management were men and its culture was masculine in style, particularly emphasising an interest in engineering and technical mastery (Faulkner, 2000). Indeed there was much interest, pleasure, and admiration for elegant products of engineering (Hacker, 1991). When one of their fighter planes flew over (an event difficult to ignore on account of the engine noise), offices would clear as employees went outside to admire the display of a beautiful machine. A certain amount of military terminology was used, sometimes ironically, in day-to-day work. A number of employees had links with the armed forces. MAC was exclusively involved in the defence industry, with the UK’s MoD being its largest customer and other approved governments buying its products.

As a manufacturing company in an economy where manufacturing was in steep decline and with its ties to the defence industry, if a major defence contract went elsewhere, jobs would be on the line. Despite the ‘hi-tech’ nature of its work, MAC had a traditional feel to it. The company had existed, under one name or another, right from the beginning of the avionics industry. The defence industry, and within that the defence aerospace industry, faced uncertain times as the UK government was redefining its expectations of the defence industry in post-Cold War times. It quickly came to expect much clearer demonstrations of value for money (Trim, 2001). Therefore, the ‘peace dividend’ brought about by the end of the Cold War meant uncertain times for the defence aerospace industry as military
spending was reduced significantly (Sillers & Kleiner, 1997). Yet, as an industry contributing huge amounts to the UK economy (around £5 billion per annum in export earnings Trim (2001, p. 227)), the defence industry is hugely important in terms of revenue and employment. Defence industries have civil wings (which was the case with MAC) and it was seen as important that the defence side of the business did not interfere with civil businesses. For instance, BAE Systems is a partner in a European consortium and was pledged £530 million as a government loan to develop the A3XXX aircraft to rival the USA’s Boeing 747 (Trim, 2001, p. 228).

Although not strictly a public sector organisation itself, its location in the defence industry put MAC’s business in the public sector. However, in the UK, views of public sector management were undergoing rapid change in the mid 1990s and it was seen as no longer acceptable that the taxpayer should underwrite investment (Trim, 2001). Such firms were required to be more competitive and to be held more accountable financially. Hence, quality management and value for money were becoming key concepts in the management repertoire of the UK defence industry from the mid 1990s onwards. As we discuss in the preceding section, this was at the height of the UK MoD’s interest in formal approaches to the production of software. In a climate where post-Cold War defence projects were likely to demand a shorter lead time, there was considerable interest in speeding up the software development process.

Computer technology and related activity clearly played a central role in MAC. One division of MAC, the Technical Directorate (TD), developed most of the airborne software (much of it real-time). This software clearly has a central role in ensuring aircraft performance and safety. Around 100 people were involved in developing systems computing software. It was in this division that Software Development System (SDS), a safety-critical airborne software system for flying military aircraft, was developed.

Research Methodology

The methodological approach of the research was based on action research (Myers & Avison, 2002). As several successful participant observation studies in technology based organisations have been reported in the literature (Forsythe, 2001; Low & Woolgar, 1993; Latour & Woolgar, 1979), an ethnographic approach holds much appeal. However, a strict ethnographic approach was neither feasible nor desirable in this study. As someone with technical expertise, the researcher could not claim to be the sociologist or anthropologist, more typical of reported ethnographic studies of technological systems (Low & Woolgar, 1993; Forsythe, 2001). This also meant that he was not ‘fobbed off’ by being directed into areas that the participants thought he wanted to look at or where they thought he should be interested in as happened in the Low and Woolgar (1993) case study. Based in the Quality Assurance Division (QAD) in the SQA team, early in his research, the researcher proved his technical credentials by helping run a workshop on software metrics and this helped to gain him full inclusion in the technical work. Although as a technical researcher, rather than a social researcher, it was arguably difficult for him to maintain the ‘anthropological strangeness’ which ethnographers look for in explaining the common sense and every day logistics of working life. In any case, he had been invited, through this research, to make a contribution to the improvement of SQA procedures. Therefore the research can be characterised as a form of action research (Baskerville & Wood-Harper, 1996), where potential improvements to SQA were to be seen as the learning part of the action research cycle.

Although action research receives a mixed press from the IS research community (Baskerville & Wood-Harper, 1996; Lau, 1999), it is nevertheless seen as a way of coming to grips with complex social settings where interactions with information technologies must be understood within the context of the whole organisation.
Baskerville (1999) notes the growing interest in action research methods in information systems research. Two key assumptions are that complex social settings cannot be reduced for meaningful study and that action brings understanding (Baskerville, 1999). The culture of MAC was extremely complex, as we characterise above and discuss again in what follows. Arguably, key elements would be lost were the researcher to have adopted a more distant role, relying on interviews and questionnaires rather than becoming fully immersed and contributing to the detail of the project. The researcher adopted an interpretivist approach, looking to the interpretations of the other participants of the research. But by allowing for social intervention he became part of the study, producing shared subjective meanings between researcher and subjects as coparticipants in the research (Baskerville, 1999).

For a period of over one year out of the three that the whole project lasted, the researcher spent, on average, one day per week working with MAC staff with access to a variety of staff across the organisation, and was therefore able to participate in a range of meetings and workshops and to gain a familiarity with the individuals concerned. This could not easily have been gained from interviews or surveys. These events included meetings where software quality staff considered quality policy, such as the implication of international standards, to broader meetings where technical staff were considering development methods in detail. Free access was allowed to relevant policy and development documents. This permitted an overview of the detailed practices and culture of this large and complex organisation.

**Analysis of Case Study Findings**

The initial remit of the researcher was to work with staff to optimise the use of software quality assurance within the organisation. The use of cost benefit analysis was originally suggested by senior management. Given our characterisation of the UK defence industry’s particular focus on management of quality and value for money, as described above, it is entirely in keeping with the industry’s changing needs that the researcher was initially directed into these areas. The researcher viewed it as problematic to assign monetary cost to SQA activities, and even harder to assign monetary benefits. However, these concerns were never addressed directly in the project as it soon emerged that there was greater interest in a new approach to software development being pioneered by MAC.

Ince (1994, p. 2-3) tells the story of a junior programmer’s first day in a new job. A senior programmer shows him around, advising him where to buy the best sandwiches at lunchtime, where to find the best beer after work, and other similarly important matters. Then the senior colleague points to a door. ‘Whatever you do don’t go through that door, the people there have been given the job of stifling our creativity.’ The door, of course, led to the quality assurance department.

The staff of MAC’s Quality Assurance Division expressed some similar feelings, albeit less dramatically. They wanted to act as consultants, offering a measure of creativity to the technical development process, although safely wrapped in appropriate quality assurance processes, but all too often they felt like the police. The strong awareness of the safety-critical nature of software development, and the related fairly advanced organisation of quality assurance in MAC, thanks in no small measure to the necessity to adhere to MoD standards, meant that SQA was never going to get quite the negative press that it attracted in Ince’s (1994) anecdote. Nevertheless, there was still some feeling that the Quality Assurance Division could be brought on board in a project some time after the Technical Division had time to do the creative part.

Hence, TD had been prototyping the new SDS system for about a year when they decided to bring in Quality Assurance Division. As we explain
below, the newness of the style of development in SDS made it unclear how it was to be quality assured. Unsure of how to proceed, the SQA manager turned to the researcher for suggestions. The researcher now became involved in investigating the use of the new software development approach, which would involve the inclusion of computer generated program code (’auto-code’) in safety-critical airborne software systems, leading to the approval of the new approach and its incorporation into MAC’s software quality assurance systems.

Although there has been a long tradition of using computers to aid the process of software engineering itself, such CASE tools (Pressman, 2005) have not generally been used to generate safety-critical code (this was always written by human programmers). The new MAC SDS was an ambitious system whose targets were principally to reduce avionics systems development time by 40% and the cost by 30%, whilst maintaining the very high quality standards necessary for computer-based system which fly—and therefore can crash—military aircraft.

A key aspect of SDS was process integration using an integrated modeling environment. There was consequentially a heavy reliance on automated methods. A specification was developed in a formal modeling language and this generated programming code automatically. In particular, automatic code generation was eventually to lead to aircraft flying ‘auto-code’ in safety-critical systems. Two aspects of SDS stand out in the climate of defence spending of the mid 1990s.

First, there was pressure to reduce costs and show value for money. Second, the use of formal methods in computer programming received a huge boost in the mid-1990s through the Defence standard DEF Stan 00-55 which mandated the use of formal methods base approaches in safety-critical software. It is not surprising that there was considerable interest in a system which offered the promise of considerably reduced software production times.

MAC invested a great deal of money and time in SDS in the hope that the improved time-scales which SDS promised, together with reduced costs, could keep major current aircraft developments on course. This was particularly important in an environment of political intervention and considerable public interest and concern over escalating costs and delivery times in the public sector, including the defence industry. These benefits could only accrue to MAC if the quality, that is, correctness of the software, could be assured.

SDS was heavily dependent on software (CASE) tools. MAC had used these for many years, and had procedures in place for their qualification (i.e., acceptance) in certain circumstances. However, these applied to mission-critical rather than safety-critical systems. Furthermore, the movement towards auto-generated code led to a different environment than one where tools improved and speeded up the design process, but where failure would show up and be merely time-wasting. There was seen to be a need for a major improvement/update of these procedures, a quantum change, before they would be acceptable for safety-critical applications.

Some tools being used had major world-wide user communities, associated academic conferences, and came from supposedly secure and reliable suppliers. Others might not be so well supported, both intellectually and commercially. (For instance, it might be no use having an ideal tool if the supplier was small and unlikely to survive for many years.) Methods already existed for supplier qualification. These methods were undertaken by software quality staff. However, the qualification of these suppliers could be a crucial issue in the qualification of the tool and ultimately the integrity of the avionics system. The issue was not merely one of qualification, it was also one of demonstration of qualification to customers. Ultimately, the need in some sense to prove the new methods became paramount. Hence we can see that quality procedures did not just involve procedures, such as code walkthroughs through
which software teams could persuade themselves to trust program code, they also applied to the question of choosing and trusting suppliers.

A number of meetings took place with members of the SDS team. This discussion was very useful for an understanding of SDS and gave the researcher a richer understanding of the SQA needs. It soon became apparent that the necessary fundamental problems with SQA in SDS were going to be difficult to answer.

The difficulties were centred around two conflicting ideas. The first of these was that for the persuasive rational argument approach to be successful there would be a need for a group of professionals to participate in code walkthroughs, with consequent discussion and persuasion. On the face of it, this was simply not possible, since the computer which wrote the auto-code could not take part in such a discussion. Alternative approaches were considered. Clearly there would be a stage before the auto-code (at the requirements specification level) where human agents were involved, but this was found to be too high level to meet the relevant military standards (the US MIL-STD-498 [1994] and the UK standard 00-55 [MoD, 1997]). Both standards are very specific about the exact conduct of the necessary walkthrough. It had to be a code walkthrough.

On the other hand, for the formal proof approach method to work, there would first need to be such a formal proof. This did not seem within the capability of the QAD itself, despite the division being quite well resourced. MAC referred back to the auto-code tools suppliers, but once again there was no such proof and no realistic possibility of achieving such a proof. Although MAC was an important customer for the auto-code tool suppliers, they were not prepared to expend the necessary resources. Furthermore, a ‘weakest link’ argument demonstrates a fundamental flaw with the formal approach in computer systems. If the auto-code tool itself could be formally verified, it would then become necessary also to consider the operating system on which the tool would run and the hardware systems involved. Potentially this could involve a seemingly infinite regression of hardware and software systems having to be proved correct, where the system is only as good as its weakest link. Frustration grew as no solution was forthcoming and ultimately SDS was shelved indefinitely.

We have argued that mathematical proof is essentially a human achievement between members of the expert mathematical community who are persuaded of the correctness of mathematical proofs because they trust each other. These processes of trust are replicated in the procedures that have been developed in software engineering, and within that, software quality assurance. As part of the defence industry, developing safety-critical systems, MAC had highly developed SQA procedures which were obliged to follow international military standards. Their code walkthroughs, which are analogous to the ways mathematicians achieve trust in a proof, were an important part of such quality procedures. Formal methods offer the promise of an attractive certainty and control over software production and hence control over the work processes of human programmers. They also offer the promise of automatic verification of software systems which, potentially, could be much cheaper than traditional human based approaches to the verification of software through traditional SQA procedures.

SDS achieved very little despite the huge efforts put into it by the many people working for MAC. Although it was not, at the time, formulated in such stark terms, success was elusive because an attempt was being made to achieve the impossible: namely using auto-code whilst being held to quality assurance procedures which demanded code walkthroughs which could not possibly be achieved in an auto-code system. Attempts were made to consider formally proving the correctness of the auto-code. In addition to supplier reluctance, this raised the spectre of the infinite regress. If
one looks to proving the auto-code correct, then
the operating system must be proved correct, the
hardware platform and so on.

This was at the height of interest in formal
methods for safety-critical systems for defence,
a view embodied in Def Stan 00-55. The rise
of formal methods is crucially linked to the de-
fence industry. The interest in formal methods
and automated approaches arrived as pressure
mounted on Western governments to prove cost
effectiveness due to the changing nature of defence
developments after the end of the Cold War and
the need to avoid litigation for software that might
be implicated in an accident. Yet the difficulties of
applying formal methods in systems of any level
of complexity and the need to trust the program
code acted as a spur to maintain complex human
centred software quality assurance procedures.

CONCLUSION:
TRUSTING COMPUTERS

There is much evidence that we already do trust
computers in many walks of life without formal
proof or other formal demonstration, even to the
extent of trusting safety-critical systems such as
the ‘fly by wire’ software in the Boeing 777
airliner, two million lines of code which have not
been fully proved (Lytz, 1995). Expert mathemati-
cians have begun to accept computer generated
proofs, albeit in qualified ways (Chang, 2004).
entrepreneurs’ of computerised risk ensure that
warnings about computerised risk are heeded
so that safety-critical software is avoided and,
where it is unavoidable, much care is taken over
its development. Military standards, so detailed
about the use of formal methods in software design
and attendant work processes in the 1990s, have
moved a decade later to be much less prescriptive
about the work methods of ensuring software
quality, thereby allowing for the crucial element
of human inspection in order that the software
may be trusted. As Collins (1990) notes, we are
remarkably accommodating to computers, mak-
sing sense of them and involving them in our social
networks, and will continue to find imaginative
ways of doing so. This echoes Nissenbaum’s (1999)
view that we may trust computers if we can treat
them as ‘agents.’ We may meaningfully ascribe
intentions and reasons to them.

In this article we have sought to tell a story of
trust, in particular how software may be trusted
when it is not produced by a human programmer.
This involves consideration of a complex
set of discourses including the question of math-
ematical proof and how proof is achieved within
mathematical communities. We see a similar
need to replicate such human processes of trust
in trusting computer systems. We have argued
that the making of standards to be applied within
software quality assurance procedures shows
ways in which mechanisms of trust are inscribed
in software standards. Our case study, an action
research project in a military aircraft company,
demonstrates the difficulties which occur when
quality assurance procedures involving code
walkthroughs—procedures with built-in human
trust mechanisms—are incommensurable with
a system which relies on auto-code. The climate
of defence research and spending was a major
influence, both on our case study and the wider
development of standards. There is a continued
tension between needing to trust and trying to
control: trusting the software and controlling its
production. The story which we tell here is one
of continuing human ingenuity in finding ways
of trusting computer software.

REFERENCES

objects. In W. E. Bijker & J. Law (Eds),
Shaping technology/building society: Studies in


Davies, B. (2006, October 3). Full proof? Let’s trust it to the black box. Times higher education supplement.


Chapter 7.17
Global Information Ethics: The Importance of Being Environmentally Earnest

Luciano Floridi
Universita degli Studi di Bari, Italy, & Oxford University, UK

ABSTRACT

The article argues that Information Ethics (IE) can provide a successful approach for coping with the challenges posed by our increasingly globalized reality. After a brief review of some of the most fundamental transformations brought about by the phenomenon of globalization, the article distinguishes between two ways of understanding Global Information Ethics, as an ethics of global communication or as a global-information ethics. It is then argued that cross-cultural, successful interactions among micro and macro agents call for a high level of successful communication, that the latter requires a shared ontology friendly towards the implementation of moral actions, and that this is provided by IE. There follows a brief account of IE and of the ontic trust, the hypothetical pact between all agents and patients presupposed by IE.

INTRODUCTION:
FROM GLOBALIZATION TO INFORMATION ETHICS

Globalization is a phenomenon too complex even to sketch in this brief introduction. So I hope that I shall be forgiven if I am rather casual about many features that would deserve full attention in another context. Here, I wish to highlight just six key transformations characterising the processes of globalization. I shall label them contraction, expansion, porosity, hybridization, synchronization, and correlation. They provide the essential background for making sense of the thesis developed in the rest of the article, which is that Information Ethics (IE) can provide a successful approach for coping with the challenges posed by our increasingly globalized reality.
Global Information Ethics

Contraction

The world has gone through alternating stages of globalization, growing and shrinking, for as long as humanity can remember. Here is a reminder:

In some respects the world economy was more integrated in the late 19th century than it is today. ... Capital markets, too, were well integrated. Only in the past few years, indeed, have international capital flows, relative to the size of the world economy, recovered to the levels of the few decades before the first world war. (The Economist, 1997)

The truth is that, after each “globalization backlash” (think of the end of the Roman or British Empires), the world never really went back to its previous state. Rather, by moving two steps forward and one step back, sometime towards the end of the last century the process of globalization reached a point of no return. Today, revolutions or the collapse of empires can never shrink the world again, short of the complete unravelling of human life as we know it. Globalization is here to stay.

Globalization has become irreversible mainly thanks to radical changes in worldwide transport and communications (Brandt & Henning, 2002). Atoms and bytes have been moving increasingly rapidly, frequently, cheaply, reliably, and widely for the past 50 years or so. This dramatic acceleration has shortened the time required for any interactions: economic exchanges, financial transactions, social relations, information flows, movements of people, and so forth (Hodel, Holderegger & Lüthi, 1998). And this acceleration has meant a more condensed life and a contracted physical space. Ours is a smaller world, in which one may multitask fast enough to give and have the impression of leading parallel lives. We may regain a nineteenth-century sense of distance (space) and duration (time) only if one day we travel to Mars.

Expansion

Human space in the twenty-first century has not merely shrunk, though. ICTs have also created a new digital environment, which is constantly expanding and becoming progressively more diverse. Again, the origins of this global, transnational common space are old. They are to be found in the invention of recording and communication technologies that range from the alphabet to printing, from photography to television. But it is only in the last few decades that we have witnessed a vast and steady migration of human life to the other side of the screen. When you ask, “Where were you?,” it is now normal and common to receive the answer “Online”. More than 6 million people throughout the world play World of Warcraft, currently the leading subscription-based MMORPG (massively multiplayer online role-playing game, http://www.blizzard.com/press/060119.shtml). Globalization also means the emergence of this sort of single virtual space, sharable in principle by anyone, any time, anywhere.

Porosity

An important relation between our contracting physical space and our expanding, virtual environment is that of porosity. Imagine living as a flat figure on the surface of an endless cylinder. You could travel on the surface of the cylinder as a two-dimensional space, but not through it. So in order to reach any other point on the cylinder, the best you could do would be to follow the shortest path (geodesic) on the cylindrical surface. The empty space inside the cylinder would be inconceivable, as a third dimension would. Imagine now that the surface became porous and hence that a third dimension were added. The geodesics would be revolutionized, for you could travel through the vacuum encircled by the cylinder and reach the other side, thus significantly shortening your journeys. To use the rather apt vocabulary of surfing, you would be tubing: space would be curling
over you, forming a “tube”, with you inside the cylindrical space. From a 2D perspective, you would literally come in and out of space. This sort of porosity characterizes the relation now between physical and virtual space. It is difficult to say where one is when one is “tubing”, but we know that we can travel through cyberspace to interact with other physical places in a way that would have been inconceivable only a few decades ago. Telepresence (Floridi, 2005) in our porous environment is an ordinary experience and this is also what globalization means.

**Hybridization**

During the last decade or so, we have become accustomed to conceptualize our life online as a mixture between an evolutionary adaptation of analogue/carbon-based agents to a digital/silicon-based environment, and a form of postmodern, neocolonization of the latter by the former. This is probably a mistake. The threshold between analogue-carbon-offline-here and digital-silicon-online-there is fast becoming blurred, but this is as much to the advantage of the latter as it is of the former. Adapting Horace’s famous phrase2, “captive cyberspace is conquering its victor”. ICTs are as much re-ontologising (that is, modifying the essential nature of) our world as they are creating new realities. The digital is spilling over into the analogue and merging with it. This recent phenomenon is variously known as “ubiquitous computing”, “ambient intelligence”, or “the Internet of things” (ITU report, November 2005, http://www.itu.int/internetofthings), and it is, or will soon be, the next stage in the digital revolution. In the (fast approaching) future, objects will be *I*entities able to learn, advise, and communicate with each other. “RoboticCookware” is already available (http://www.vitacraft.com.nyud.net:8090/rfiq/home.html); MP3 players will soon be able to recommend new music to their users by learning from the tunes they (the users, we had better be clear) enjoyed (http://www.semanticaudio.com/). Your next fridge (http://www.lginternetfamily.co.uk/homenetwork.asp) will inherit from the previous one your tastes and wishes, just as your new laptop can import your favourite settings from the old one; and it will interact with your new way of cooking and with the supermarket Web site, just as your laptop can talk to a printer or to another computer. We have all known this in theory for some time; the difference is that it is now actually happening in our kitchen.

Globalization also means the emergence of this common, fully interactive, and responsive environment of wireless, pervasive, distributed, *a2a* (anything to anything) information processes, that works *a4a* (anywhere for any time), in real time. We are probably the last generation to experience a clear difference between onlife and online.

**Synchronization**

In a world in which information and material flows are becoming so tightly integrated and enmeshed, it is not surprising to see global patterns emerging not only from well-orchestrated operations (consider the tedious experience of any launch of a major blockbuster, with interviews in magazines, discussions on TV programs, advertisements of merchandise, and by-products throughout the world, special food products in supermarkets and fast-food, etc.), but also inadvertently, as the result of the accidental synchronization of otherwise chaotic trends.

All of a sudden, the world reads the same novel, or wears the same kind of trousers, or listens to the same music, or eats the same sort of food, or is concerned about the same problems, or cherishes the same news, or is convinced that it has the same disease. Some of this need not be the effect of any plan by some Big Brother, a secret agency, a powerful multinational or any other *deus ex machina* that is scheming behind the curtains. After all, worldwide attention span is very limited
and flimsy, and it is very hard to compete for it. The truth is that at least some global trends may merely arise from the constructive interference of waves of information that accidentally come into phase, and hence reinforce each other to the point of becoming global, through the casual and entirely contingent interaction of chaotic forces. It may happen with the stock markets or the fashion industry or dietary trends. The recurrent emergence of temporarily synchronized patterns of human behaviour, both transculturally and transnationally, is a clear sign of globalization, but not necessarily of masterminded organization. There is no intelligent plan, evil intention, autonomy, or purposeful organization in the billion snow flakes that become an avalanche. Social group behaviour is acquiring a global meaning. The distributed power that generates Wikipedia is the other side of the dark, mindless stupidity of millions of slaves of fashions and trends.

**Correlation**

Imagine a safety net, like the one used in a circus. If it is sufficiently tight and robust, the heavier the object that falls into it, the larger the area of the net that will be stretched, sending waves of vibration throughout the net. Globalization also refers to the emergence of a comparable net of correlations among agents all over the world, which is becoming so tight and sensitive that the time lag in the transmission of the effects of an event “dropping” on it is fast shortening, to the point that sometimes there is almost no distinction between what counts as local or remote. Global often means not everywhere but actually delocalized, and in a delocalized environment social friction is inevitable, as there is no more room for agents that allows for absorption of the effects of their decisions and actions. If anyone moves, the global boat rocks.

**Globalising Ethics**

If we consider now the profound transformations just sketched, it would be rather surprising if they did not have serious implications for our moral lives (see Ess, 2002; Weckert, 2001). In a reality that is more and more physically contracted, virtually expanded, porous, hybridized, synchronized, and correlated, the very nature of moral interactions, and hence of their ethical analysis, is significantly altered. Innovative forms of agenthood are becoming possible; new values are developing and old ones are being reshaped; cultural and moral assumptions are ever more likely to come into contact when not into conflict; the very concepts of what constitutes our “natural” environment and our enhanced features as a biological species are changing; and unprecedented ethical challenges have arisen (a reference to the notorious problem of privacy is de rigueur here), just to mention some macroscopic transformations in which globalization factors, as sketched above, play an important role.

What sort of ethical reflection can help us to cope successfully with a world that is undergoing such dramatic changes? Local approaches are as satisfactory as burying one’s head in home values and traditions. The ethical discourse appears to be in need of an upgrade to cope with a globalized world. Each ethical theory is called upon to justify its worldwide and cross-cultural suitability. This seems even more so if the theory in question seeks to address explicitly the new moral issues that arise from the digital revolution, as it is the case with IE.

I shall say more about IE in the next two sections. The specific question that I wish to address is whether, in a world that is fast becoming more and more globalized, information ethics can provide a successful approach for dealing with its new challenges. I shall argue in favour of a positive answer. But to make my case, let me first clarify what global information ethics may mean.
Global-Communication Ethics vs. Global-Information Ethics

There are at least two ways of understanding Global Information Ethics: as an ethics of global communication (Smith, 2002) or as a global-information ethics (Bynum & Rogerson, 1996). Since I shall concentrate only on the latter, let me briefly comment on the former first.

Global-information ethics, understood as an ethics of worldwide communication, may be seen as a commendable effort to foster all those informational conditions that facilitate participation, dialogue, negotiation, and consensus-building practices among people, across cultures and through generations. It is an approach concerned with new and old problems, caused or exacerbated by global communications or affecting the flow of information. Global-communication ethics is therefore a continuation of policy by other means, and it does not have to be reduced to a mere gesture towards the importance of mutual respect and understanding (meeting people and talking to each other can hardly do any harm and often helps). It is, however, faced by the serious problem of providing its own justification. What sort of ethical principles of communication and information are to be privileged and why? Is there any macroethics (e.g., some form of consequentialism or deontologism or contractualism) that can rationally buttress a global-communication ethics? And is not any attempt at providing such a macroethics just another instance of “globalization” of some values and principles to the disadvantage of others? Without decent theorization, the risk is that we will reduce goodness to goodiness and transform the ethical discourse into some generic, well-meant sermon. At the same time, a robust foundation for a global-communication ethics may easily incur the problem of failing to respect and appreciate a plurality of diverse positions. The dilemma often seems to be left untouched, even when it is not overlooked. The good news is that it may be possible to overcome it by grounding a global-communication ethics on a global-information ethics.

Global-Information Ethics and the Problem of the Lion

If we look at the roots of the problem, it seems that:

1. In an increasingly globalized world, successful interactions among micro and macro agents belonging to different cultures call for a high level of successful communication; but
2. Successful, cross-cultural communications among agents require, in their turn, not only the classic three “e”s—embodiment, embeddedness and hence experience (a sense of “us-here-now”)—but also a shared ontology (more on this presently); and yet
3. Imposing a uniform ontology on all agents only seems to aggravate the problem, globalization becoming synonymous with ontological imperialism.

By “ontology” I do not mean to refer here to any metaphysical theory of being, of what there is or there is not, of why there is what there is, or of the ultimate nature of reality in itself. All this would require a form of epistemological realism (some confidence in some privileged access to the essential nature of things) that I do not hold, and that, fortunately, is not necessary to make my case. Rather, I am using “ontology” to cover the outcome of a variety of processes that allow an agent to appropriate (be successfully embedded in), semanticize (give meaning to and make sense of), and conceptualize (order, understand, and explain) the agent’s environment. In simplified terms, one’s ontology is one’s world; that is, the world as it appears to, is experienced and interacted with, by the agent in question.
Agents can talk to each others only if they can partake to some degree in a shared ontology anchored to a common reality to which they can all refer.4

Imagine two solipsistic minds, α and β, disembodied, unembedded, and devoid of any experience. Suppose them living in two entirely different universes. Even if α and β could telepathically exchange their data, they could still not communicate with each other, for there would be absolutely nothing that would allow the receiver to interpret the sender. In fact, it would not even be clear whether any message was being exchanged at all.

The impossibility of communication between α and β is what Wittgenstein (2001) had in mind, I take it, when he wrote that “if a lion could talk, we could not understand him.” The statement is obviously false (because we share with lions a similar form of embeddedness and embodiment, and hence experiences like hunger or pain) if one fails to realize that the lion is only a placeholder to indicate an agent utterly and radically different from us, like our α and β. The lion is a Martian, someone you simply cannot talk to because it is “from another ontology”.5

From this perspective, the famous phrase hic sunt leones (here there are lions) acquires a new meaning. The phrase occurred on Roman maps to indicate unknown and unexplored regions beyond the southern African borders of the empire.6 In a Wittgensteinian sense, the Romans were mapping the threshold beyond which no further communication was possible at all. They were drawing the limits of their ontology. What was beyond the border, the locus inhabited by the lions, was nothing, a nonplace. Globalization has often meant that what is not inglobate simply is not, that is, fails to exist.

We can now formulate the difficulty confronting a global-information ethics as the problem of the lion: cross-cultural communication, which is the necessary condition for any further moral interaction, is possible only if the interlocutors partake in a common ontology. When Crusoe and Friday meet, after 25 years of Crusoe’s solitude, they can begin to communicate with each other only because they share the most basic ontology of life and death, food and shelter, fear and safety. Agents may be strangers to each other (“stranger” being an indexical qualification7). They do not have to speak the same language, empathize, or sympathize. But they do need to share at least some basic appropriation, semanticization, and conceptualization of their common environment, as a minimal condition for the possibility of any further moral interaction.

Can information ethics provide a solution to the problem of the lion? The short answer is yes; the long one is more complicated and requires a brief diversion, since it is now necessary to be more explicit about what I mean by information ethics.

Global Information-Ethics and Its Advantages

Information ethics8 is an ontocentric, patient-oriented, ecological macroethics. An intuitive way to unpack this definition is by comparing IE to other environmental approaches.

Biocentric ethics usually grounds its analysis of the moral standing of bio-entities and eco-systems on the intrinsic worthiness of life and the intrinsically negative value of suffering. It seeks to develop a patient-oriented ethics in which the “patient” may be not only a human being, but also any form of life. Indeed, land ethics extends the concept of patient to any component of the environment, thus coming close to the approach defended by information ethics. Any form of life is deemed to enjoy some essential proprieties or moral interests that deserve and demand to be respected, at least minimally if not absolutely, that is, in a possibly overridable sense, when contrasted to other interests. So biocentric ethics argues that the nature and well-being of the patient of any action constitute (at least partly) its moral standing and that the latter makes important claims on the interacting agent, claims that in principle ought to contribute to guiding the agent’s ethical decisions and constraining the agent’s moral
behaviour. The “receiver” of the action is placed at the core of the ethical discourse, as a centre of moral concern, while the “transmitter” of any moral action is moved to its periphery.

Now substitute “existence” for “life” and it should become clear what IE amounts to. IE is an ecological ethics that replaces biocentrism with ontocentrism. It suggests that there is something even more elemental than life, namely being—that is, the existence and flourishing of all entities and their global environment—and something more fundamental than suffering, namely entropy. The latter is most emphatically not the physicists’ concept of thermodynamic entropy. Entropy here refers to any kind of destruction or corruption of entities understood as informational objects (not as semantic information, take note), that is, any form of impoverishment of being, including nothingness, to phrase it more metaphysically.9

We are now ready to appreciate some of the main advantages offered by information ethics when it comes to the new challenges posed by globalization.

1. Embracing the New Informational Ontology

Not only do we live in a world that is moving towards a common informational ontology, we also experience our environment and talk and make sense of our experiences in increasingly informational ways. Information is the medium. This calls for an ethics, like IE, that, by prioritising an informational ontology, may provide a valuable approach to decoding current moral phenomena and orienting our choices.

2. Sharing a Minimal, Horizontal, Lite Ontology

There is a risk, by adopting an ontocentric perspective, as IE suggests, that one may be merely exchanging one form of “centrism” (American, Athenian, Bio, European, Greek, Male, Western, you-name-it) with just another, perhaps inadvertently, thus failing to acknowledge the ultimate complexity, diversity, and fragility of the multicultural, ethical landscape with which one is interacting. We saw how the problem of the lion may become a dilemma. This justified concern, however, does not apply here because IE advocates a minimal informational ontology, which is not only timely, as we have just seen, but also tolerant of, and interfaceable with, other local ontologies. Thick cultures with robust, vertical ontologies—that is, deeply-seated, often irreconcilable, fundamental conceptions about human nature, the value and meaning of life, the nature of the universe and our place in it, society and its fair organization, religious beliefs, and so forth—can more easily interact with each other if they can share a lite, horizontal ontology as little committed to any particular Weltanschaung as possible. The identification of an absolute, ultimate, monistic ontology, capable of making all other ontologies merge, is just a myth, and a violent one at that. Yet the alternative cannot be some form of relativism. This is no longer sustainable in a globalized world in which choices, actions, and events are delocalized. There simply is not enough room for “minding one’s own business” in a network in which the behaviour of each node may affect the behaviour of all nodes. The approach to be pursued seems rather to be along the lines of what IE proposes: respect for and tolerance towards diversity and pluralism and identification of a minimal common ontology, which does not try to be platform independent (i.e., absolute), but cross-platform (i.e., portable).

As in Queneau’s Exercises in Style, we need to be able to appreciate both the ninety-nine variations of the same story10 and the fact that it is after all the same story that is being recounted again and again. This plurality of narratives need not turn
into a Babel of fragmented voices. It may well be a source of pluralism that enriches one’s ontology. More eyes simply see better and appreciate more angles, and a thousand languages can express semantic nuances that no global Esperanto may ever hope to grasp.

3. Informational Environmentalism

The ontocentrism supported by IE means that at least some of the weight of the ethical interpretations may be carried by (outsourced to) the informational ontology shared by the agents, not only by the different cultural or intellectual traditions (vertical ontologies) to which they may belong. Two further advantages are that all agents, whether human, artificial, social or hybrid, may be able to share the same minimal ontology and conceptual vocabulary; and then that any agent may take into account ecological concerns that are not limited to the biosphere.

4. Identifying the Sources and Targets of Moral Interactions

One of the serious obstacles in sharing an ontology is often how the sources and targets of moral interactions (including communication) are identified. The concept of person or human individual, and the corresponding features that are considered essential to his or her definition, might be central in some ontologies, marginal in others, and different in most. IE may help foster communication and fruitful interactions among different, thick, vertical ontologies by approaching the problem with conceptual tools that are less precommitted. For when IE speaks of agents and patients, these are neutral elements in the ethical analysis that different cultures or macro-ethics may be able to appropriate, enrich, and make more complex, depending on their conceptual requirements and orientations. It is like having an ontology of agency that is open source, and that anyone can adapt to its own proprietary Weltanshaung.

The Cost of a Global-Information Ethics: Postulating the Ontic Trust

It would be silly to conclude at this point that a global-information ethics may provide an answer to any challenge posed by the various phenomena of globalization. This would be impossible. Of course, there will be many issues and difficulties that will require substantial extensions and adaptations of IE, of its methodology and of its principles. All I have tried to do is to convince the reader that such a great effort to apply IE as a global ethics would be fruitful and hence worth making.

It would be equally wrong to assume that the adoption of IE as a fruitful approach to global challenges may come at no conceptual cost. Every ethical approach requires some concession on the part of those who decide to share it and IE is no exception.

The cost imposed by IE is summarizable in terms of the postulation of what I shall define as the ontic trust binding agents and patients. A straightforward way of clarifying the concept of ontic trust is by drawing an analogy with the concept of “social contract”.

Various forms of contractualism (in ethics) and contractarianism (in political philosophy) argue that moral obligation, the duty of political obedience, or the justice of social institutions, have their roots in, and gain their support from a so-called “social contract”. This may be a real, implicit, or merely hypothetical agreement between the parties constituting a society (e.g., the people and the sovereign, the members of a community, or the individual and the state). The parties accept to agree to the terms of the contract and thus obtain some rights in exchange for some freedoms that, allegedly, they would enjoy in a hypothetical state of nature. The rights and responsibilities of the parties subscribing to the agreement are the terms of the social contract, whereas the society, state, group, an so forth, are the entity created for the purpose of enforcing the
agreement. Both rights and freedoms are not fixed and may vary, depending on the interpretation of the social contract.

Interpretations of the theory of the social contract tend to be highly (and often unknowingly) anthropocentric (the focus is only on human rational agents) and stress the coercive nature of the agreement. These two aspects are not characteristic of the concept of ontic trust, but the basic idea of a fundamental agreement between parties as a foundation of moral interactions is sensible. In the case of the ontic trust, it is transformed into a primeval, entirely hypothetical \textit{pact}, logically pre-dating the social contract, which all agents cannot but sign when they come into existence, and that is constantly renewed in successive generations.\textsuperscript{11}

The sort of pact in question can be understood more precisely in terms of an actual trust. Generally speaking, a trust in the English legal system is an entity in which someone (the trustee) holds and manages the former assets of a person (the trustor, or donor) for the benefit of certain persons or entities (the beneficiaries). Strictly speaking, nobody owns the assets. Since the trustor has donated them, the trustee has only legal ownership and the beneficiary has only equitable ownership. Now, the logical form of this sort of agreement can be used to model the ontic trust, in the following way:

- The assets or “corpus” is represented by the world, including all existing agents and patients;
- The donors are all past and current \textit{generations} of agents;
- the trustees are all current \textit{individual} agents;
- The beneficiaries are all current and future \textit{individual} agents and patients.

By coming into being, an agent is made possible thanks to the existence of other entities. It \textit{is} therefore bound to all that already is both \textit{unwillingly} and \textit{inescapably}. It \textit{should} be so also \textit{caringly}. Unwillingly, because no agent wills itself into existence, though every agent can, in theory, will itself out of it. Inescapably, because the ontic bond may be broken by an agent only at the cost of ceasing to exist as an agent. Moral life does not begin with an act of freedom but it may end with one. \textit{Caringly} because participation in reality by any entity, including an agent—that is, the fact that any entity is an expression of what exists—provides a right to existence and an invitation (not a duty) to respect and take care of other entities. The pact then involves no coercion, but a mutual relation of appreciation, gratitude, and care, which is fostered by the recognition of the dependence of all entities on each other. A simple example may help to clarify further the meaning of the ontic trust.

Existence begins with a gift, even if possibly an unwanted one. A foetus will be initially only a beneficiary of the world. Once it is born and has become a full moral agent, it will be, as an individual, both a beneficiary and a trustee of the world. It will be in charge of taking care of the world, and, insofar as it is a member of the generation of living agents, it will also be a donor of the world. Once dead, it will leave the world to other agents after it and thus becomes a member of the generation of donors. In short, the life of an agent becomes a journey from being only a beneficiary to being only a donor, passing through the stage of being a responsible trustee of the world. We begin our career of moral agents as strangers to the world; we should end it as friends of the world.

The obligations and responsibilities imposed by the ontic trust will vary depending on circumstances but, fundamentally, the expectation is that actions will be taken or avoided in view of the welfare of the whole world.

The ontic trust is what is postulated by the approach supported by IE. According to IE, the ethical discourse concerns any entity, understood informationally, that is, not only all persons, their cultivation, well-being, and social interac-
tions, not only animals, plants, and their proper natural life, but also anything that exists, from buildings and other artefacts to rivers and sand. Indeed, according to IE, nothing is too humble to deserve no respect at all. In this way, IE brings to ultimate completion the process of enlargement of the concept of what may count as a centre of a (no matter how minimal) moral claim, which now includes every instance of being understood informationally, no matter whether physically implemented or not. IE holds that every entity, as an expression of being, has a dignity, constituted by its mode of existence and essence (the collection of all the elementary proprieties that constitute it for what it is), which deserve to be respected (at least in a minimal and overridable sense) and hence place moral claims on the interacting agent and ought to contribute to guiding and constraining the agent’s ethical decisions and behaviour. The ontic trust (and the corresponding ontological equality principle among entities) means that any form of reality (any instance of information/being), simply by the fact of being what it is, enjoys a minimal, initial, overridable, equal right to exist and develop in a way which is appropriate to its nature.¹²

The acceptance of the ontic trust requires a disinterested judgement of the moral situation from an objective perspective, that is, a perspective which is as non-anthropocentric as possible. Moral behaviour is less likely without this epistemic virtue. The ontic trust is respected whenever actions are impartial, universal and “caring” towards the world.

CONCLUSION

One of the objections that is sometimes made against IE is that of being too abstract or theoretical to be of much use when human agents are confronted by very concrete and applied challenges (Siponen, 2004). Unfortunately, this is an obvious misunderstanding. Imagine someone who, being presented with the declaration of human rights, were to complain that it is too general and inapplicable to solve the ethical problems the person is facing in a specific situation, say in dealing with a particular case of cyberstalking in the company that employs the person. This would be rather out of place. The suspicion is that some impatience with conceptual explorations may betray a lack of understanding of how profound the revolution we are undergoing is, and hence how radical the rethinking of our ethical approaches and principles may need to be in order to cope with it. IE is certainly not the declaration of human rights, but it seeks to obtain a level of generality purporting to provide a foundation for more applied and case-oriented analyses. So the question is not whether IE is too abstract—good foundations for the structure one may wish to see being built inevitably lie well below the surface—but whether it will succeed in providing the robust framework within which practical issues of moral concern may be more easily identified, clarified, and solved. I agree that it is in its actual applications that IE, as a global ethics for our information society, will or will not qualify as a useful approach; yet the need to build on the foundation provided by IE is an opportunity, not an objection.

REFERENCES


Ess, C. (2002). Computer-mediated colonization, the renaissance, and educational imperatives for


ENDNOTES

1 For a very synthetic but well-balanced and informed overview, I would recommend Held and McGrew (2001). In their terminology, I am a subscriber to the transformationalist approach, according to which “globalization does not simply denote a shift in the extensity or scale of social relations and activity. Much more significantly, argue the transformationists, it also involves the spatial re-organization and re-articulation of economic, political, military and cultural power” (see Held et al., 1999).

2 *Graecia capta ferum victorem cepit*- Epistles.

3 How an ontology is achieved and what sort of philosophical analysis is required to make sense of its formation is not a relevant matter in this context, but the interested reader may wish to see Floridi and Sanders (in press).

4 More technically, this means that two agents can communicate only if they share at least some possible level of abstraction. On the method of abstraction see Floridi and Sanders (2004) and Floridi and Sanders (in press).

5 If it took endless time and efforts to decipher the hieroglyphics, imagine what sense an extraterrestrial being could make of a message in a bottle like the plaque carried by the Pioneer spacecraft (http://spaceprojects.arc.nasa.gov/Space_Projects/pioneer/PN10&11.html)
Unfortunately, we do not have African maps drawn from the “lions’ perspective”. The Da Ming Hun Yi Tu, or Amalgamated Map of the Great Ming Empire, the oldest map of Africa known so far, dates back “only” to 1389.

Indexical expressions, such as “here”, “yesterday”, or “I”, acquire their meaning or reference depending on who utters them and in which circumstances. Thus, “stranger” is indexical (people are strangers to each others), whereas the original meaning of “barbarian” is not, if we believe its Greek etymology to be “to babble confusedly”, that is, someone who is unable to speak Greek properly.

The IEG, a research group in Oxford, has developed a general interpretation of Information Ethics in a series of papers. Here I provide a summary based on Floridi [in press]. The interested reader is invited to check the Web site of the group at http://web.comlab.ox.ac.uk/oucl/research/areas/ieg/.

Destruction is to be understood as the complete annihilation of the object in question, which ceases to exist; compare this to the process of “erasing” an entity irrevocably. Corruption is to be understood as a form of pollution or depletion of some of the properties of the object, which ceases to exist as that object and begins to exist as a different object minus the properties that have been corrupted or eliminated. This may be compared to a process degrading the integrity of the object in question.

On a crowded bus, a narrator observes a young man with a long neck in a strange hat yell at another man whom he claims is deliberately jostling him whenever anyone gets on or off the bus. The young man then sits down in a vacant seat. Two hours later the same narrator sees that same young man with another friend, who is suggesting that the young man have another button put on his overcoat.

There are important and profound ways of understanding this Ur-pact religiously, especially but not only in the Judeo-Christian tradition, where the parties involved are God and Israel or humanity, and their old or new covenant (διαθήκη) makes it easier to include environmental concerns and values otherwise overlooked from the strongly anthropocentric perspective prima facie endorsed by contemporary contractualism. However, it is not my intention to endorse or even draw on such sources. I am mentioning the point here in order to shed some light both on the origins of contractualism and on a possible way of understanding the onto-centric approach advocated by IE.

In the history of philosophy, a similar view can be found advocated by Stoic and Neoplatonic philosophers, and by Spinoza.
Chapter 7.18

Emotional Digitalization as Technology of the Postmodern: A Reflexive Examination from the View of the Industry

Claus Hohmann
Bergische Universitaet Wuppertal, Germany

ABSTRACT

This article introduces emotional digitalization as a phenomenon of future information systems. It argues that emotional digitalization is a progress that will lessen the gap between technology and humanity, as well as between computer and man. The author develops and verifies his assumption besides theoretical references arising from his experiences with the information technology within the BrandLand Autostadt.

INTRODUCTION

“The inmost force/which binds the world, and guides its course” is no longer only a philosophical thought or religious voice, as in Goethe’s “Faust,” verse 382-383. But they are also not just bits and bytes, cable and monitors. The connection of these elemental substructures of modern civilization first shows where the path must go. The functional elements of our daily life must subordinate themselves to our principles of thought and aesthetics. And they must also appeal to the people in an aesthetic, as well as communicative and sense. The modern, as a synonym for freedom and democracy, could thereby obtain a new dimension — to become an ethical entity. The technology of the third millennium must also define the new standards.

As the postmodern breaks down the barriers between art and pleasure, it breaks through the wall between technology and emotion. The aesthetic sensation from looking at, for example, a corporate homepage, accessible to all, suggests a fundamental change in the meaning which will be attributed to the presentation of technology today, when in direct comparison with the functional-rationalistic views of modern, purely informational Web sites. The Autostadt, Volkswagen’s new communication platform for marketing and
Emotional Digitalization as Technology of the Postmodern

culture, is on several levels a culmination of different developments, which operate under the term “postmodern” and therefore contribute to theoretical discussion.

This radical change of the basic conditions of our life has put us on a path which will be indicated by a catchy, yet imprecise phrase: “postmodernity.” The rationalization of the modern is being counteracted by the pluralization of our culture and humanization of technology, which was demanded, but not achieved, by the modern (Giddens, 1995). With Anthony Giddens, whose critique is applied here, we regard “posthistory“ as “a succession of immanent changes,” in order to face the prevalent conceptual dilemma (Giddens, 1995). One focuses the view to information technologies and sees that the humanization has an effect, particularly on aesthetics. And here alongside the desired effects lies a formidable danger: does IT go in the same ambivalent direction as the classic technologies, and will it give the manipulators of this world a leg up?

THE WORLD TODAY AND TOMORROW

The core problems of the industrial society (saturated markets, over-indebtedness, mass unemployment, etc.) demanded drastic mental reorientation and have at the same time uncovered innovation potentials. Spirituality and mental dimensions begin to replace the abstraction — the dream becomes a legend. The new technologies of the third millennium must also surrender to these new demands. With help from the most current technologies, we learn to tell stories. Contrary to the findings of Lyotard from the end of the great narrations of history1, one could also speak about a fragmentation, which transfers this narration from people to technology, especially the digital technologies. With a new neologism, one speaks of the necessity of a “homuter society,” which is in a position to propitiate the people with (information) technologies (Haefner, 1984). Haefner uses this term freely in order to verbalize his skepticism of the possibilities of future developments, which was obtained during the Cold War. However, we could also read it with a conscious hope.

Thus, a great possibility of the future lies in the development of technologies, stimulating creativity and inspiring thoughts. For example, the aesthetic layout of an intranet and the constantly used user interfaces eases the employee’s accessing of the media, prevents fatigue from using information and inspires emotions in the exposure to IT in everyday work. The effect is to state a higher degree of identification to the corporation, a more emotional, eventually better relationship to the employer and, ultimately and ideally considered, an increased labor efficiency accompanied by increased happiness. In this manner, IT of the postmodern can manage the reconciliation of the contradictions between technology and art, which are unimaginably present in the modern.

The IT of the present has the function of supporting the thoughts and actions of the individual and not to automate their work in order to eventually replace them. Comparably, the focus should be put on the installation of technologies that will save time. IT must be used to support work processes in order to optimize the concrete procedures and reduce the time requirements. Networked, Web-based systems with access points for all employees could be a first step. Important here are the user interfaces, which are adjusted to the work processes, aesthetically configured, self-explanatory, intuitive and can be understood spontaneously, thereby helping save time and increase room for further assignments.

The same exposed meaning, which electrification and mechanization had for the modern, could be digitization and computerization in the postmodern. The unreality, which challenges our terminology and understanding in increased
Emotional Digitalization as Technology of the Postmodern

measure, seems to be a display of something new in relation to the more palpable technological advances of the modern. We see, in this transition of information technology, a movement for a new understanding of our time, which can be labelled as “postmodern,” which is in this sense, in the shadows of abstract, fragmentary, and simultaneously emotional media. Thereby, no clearly differentiated epochs or outlined theories should be suggested — it should continue to deal with a change in our perception and “Lebenswelt.” In doing so, we understand postmodern as the “conscious of a socio-historical transition” and not as a “meta-historical category” of each epoch.

One of the basic requirements of this transition is the change from a technology that replaces the people to a technology that supports the people because they “understand” it. Thereby, it is imperative to restore various elements of IT, from management decisions and business processes down to connections with trade and commerce, to an original understanding of “experience.” The anchor of the complicated, indiscernible world in tangible lives can show in this manner how postmodernity can be made visible. The metaphor can be helpful to establish the meaning behind the experience.

POSTMODERNITY AND IT

The implementation and deployment of aesthetics in the technological area is a characteristic of the realization that IT is anchored fast in our world and no longer accepted as a modern achievement of an exceptional position outside our consciously experienced everyday life. This insight allows changed societal requirements to be inferred, but what about the above meaning is postmodern? Important is our understanding after the liberation from a conditioned past in a (self-) conscious act. As the technologies of modernity stand out due to their distinctiveness and are clearly contrasted with the people and their emotional world in which they coldly, precisely, and rationalistically fulfill their duties, the IT of postmodernity goes down a different path.

With the integration of aesthetics and emotion, the digitization completes the circle and appears as a new quality. This is only successful because IT has already begun to reflexively examine and question itself and its success, and not in the least can be measured as a clearly calculated result of other factors. A basic requirement, therefore, is the development of technology that supports activities with which people are discontent as a result of the holistic pursuit of new goals, with the focus on the people. IT must be there for the people, not the other way around, and it must be aware of this. It must be noticeable, but not necessarily visible.

When the technologies in our new world function imperceptibly and no longer distort the view of the contents from the surface, new space for aesthetics and emotion is created. Technological progress can and must give the people more time, space and capacity for creativity and to unleash their imagination to make the fundamental, structural change perceptible and understandable. In sociological terms, postmodernity seems to accompany the departure from middle-class feudalism. Information technology, which has been more or less privatized by the personal computer (PC), can also lead to further democratization and contribute to free communication, but it can also lead to chaos by making corporations unmanageable.

Thus, in the end the various technology images of “two cultures,” which are “Lebenswelt”- and system-orientated, according to Charles Percy Snow, should be brought into unison. But in order to accomplish this, technology will be considered as a social phenomenon, with the help of metaphor and symbolization. The symbols must orient themselves on the life-world, in that they become a part of the life-world. Just as modernity, with
its dichotomies and ambivalences, can be said to be a failure in dialect, it is conceivable that the discussion in the postmodern, founded in the definition of difference, will reveal endeavors leading to the harmony between the various positions. Contrary to Lyotard’s assertion, unanswerable contradictions, which lie in the open availability of information, could strengthen the democratizing tendencies of information technologies.

**WITH THE EYES OF INDUSTRY**

With the beginning of the new millennium, the world changes its vision. The blind trust of technology has a conscious practice of giving way to its possibilities. For example, Volkswagen’s new corporate platform culture and marketing—the Autostadt allows this change, which is currently being experienced by industry and society, to be precisely identified. With this example, we want to discuss the topic of postmodernity from the view of information technology, in order to stimulate further thoughts from the industry. Progress is nowadays something completely different, as it questions itself. Reflexivity and self-criticism take the place of the unconscious self-awareness of modern advances. Technology along with the information technology of modernity has undertaken a rapid development, which leaves the people and their imagination helplessly behind. In the physical sense, it is completely impossible to grasp IT, and an attempt at understanding it requires shifting the focus from a purely technical, modern-rational meaning to the interpersonal: addressing emotions via aesthetics, design and a process-logical functioning.

In practical industrial terms, the example of a “gated community,” such as the Autostadt, is used here to illustrate this transformation and discuss postmodernity in the digital light of information technology, in order to stimulate further ideas. The Volkswagen Group’s Autostadt presents itself as a holistic concept, which includes all details in a complex structure: “There are no details in the execution”. The mass tourism and amusement parks shape the leisure time culture of modernity, but the Autostadt vastly transcends this. It integrates, specifically with the involvement of IT, the science of the world of industry with the goal of getting through to the people.

The privatization of technology has extended continually since the 1980s the influence and impact of IT to all areas of life, but especially to the areas of personal computers and music and television technologies, and also to all other levels of everyday life. With the Autostadt, the attempt is made to take on these societal changes and harmoniously unite them. In the Autostadt, the ubiquity of IT starts with the central purpose of the corporation, namely the delivery of new cars from the Volkswagen Group to its customers. The delivery process is completely supported by digitization. From the ordering to the handing over of the new Volkswagen to the customer, IT technologies are deployed to support the process. Already in the sales dialog, the employee, by making use of online applications, is enabled to have a virtually intimate discussion, having a positive impact.

Also presentations, such as the LupoShow (a dramatic laser show demonstrating the internal functions of an automobile via cross-section model), the SunFuelLab (a plant computer, which generates personally assigned biofuels) or the NavigationsSpiel (a virtual Labyrinth), are entirely digitally controlled. However, even internal support processes such as the lighting controls in the buildings and at events, the watering of the on-site parks and the digital menus in the restaurants are based on IT technologies and displayed systematically. In all of these areas, IT supports the employees, as it is deployed in order to assist them in effectively accomplishing their work and saving time.

Naturally, this all comes with limits: the Autostadt has not realized certain global interactivity, and their video arts suggest only a portion of the
"Iconic Turn" of these years \(^{11}\), but even with the example of the Autostadt, it’s clear that the practical IT supports a form of the philosophy of imagery. The Autostadt is thereby also active in the area of bringing to life aesthetic contents in the framework of its marketing scene, and in this lies a threshold experience.

Entities such as the Autostadt must have come into existence, not as a storm of individual thoughts, but rather as a logical consequence of the mentioned societal changes. The Autostadt is then an answer to the liberation of the postmodern individual, unique to the VW Group, and thereby not a blueprint that can be taken as a model for others. In doing so, what can the technology accomplish?

The industry serves here as an aesthetic-philosophical realization, and in the case of the Autostadt, the first deliberate BrandLand. The societal developments have encouraged, if not called for, the opening of more subtle theme parks. Others have just been, or are yet to be, opened: Daimler-Chrysler, the BMW-World and Dr. Oetker, just to name a few from the nearby surroundings. With the help of IT, the economy of aesthetics and art can be greatly improved.

The interfaces of the IT, acting as the window to foreign worlds, allow a new “cultural relocation” in the sense of “immobile mobility”\(^{12}\), which places the people back in the focus. This liberation of the realm of reality expands, in a virtual manner, the consciously experienced world of the individuals, which would be impossible without IT. The people learn, experience, fulfill desires (and generate new desires) and expand their horizons via purposefully installed, yet unconsciously experienced, IT. The IT gives the phenomenology of perception\(^{13}\) a new, different touch. Multiperspective observation possibilities allow a concrete regard to objects and forms of interactivity. In this way, IT allows us to consider something from all sides, without forcing us to change our, or its, location.

The Autostadt is also a symbol for a trend. Corporations today present themselves differently. But what exactly is the Autostadt?

THE AUTOSTADT PROJECT

At a crucial interface, residing between the company and society, and with the old factory in the background in order to illustrate the coexistence of industry and BrandLand, the Autostadt is presented as a landscaped park with integrated, individually configured brand pavilions, where nature and technology are connected in an aesthetic manner so that the relevance of the corporate values are consciously absorbed by the people via the polyvalently applied presentations. The fundamental concept does not view art as a decorative end in itself, but rather as “a link between designed environments and the visitor experiencing them”\(^{14}\).

The borders between art and presentation are blurred in this regard. Culture, understood as a variable system of seeking truth and orientation relating to a variety of artistic forms of expression, education, and living, is finally the added value, which allows the Autostadt to obtain visitors. Thereby, it shows itself as an innovative cultural platform in a new understanding of the marketing of experiences and events\(^{15}\).

In this context, great significance is given to the development of innovative technologies, as the visible, evident use of digitization would inevitably undermine the functioning of the entire concept of the Autostadt based upon art and aesthetics. The organizational level of the Autostadt has therefore been deliberately made invisible and moved to underground tunnel systems and cable ducts. This form of understatement certainly gives the nearly ambitious technology a poetic touch. And just one step further are the Web-based information
Emotional Digitalization as Technology of the Postmodern


The glass auto-towers with their transparent aesthetics give the sublimity back to the technology and astonish the observer, yet their architecture, infused with aesthetics, prevents fright from their size, which large technologies commonly provoke. And in the park, the straight line, as the enemy of everything living, is already conceptually forbidden. Intuition and emotionality operate in the place of regularization and rationalization. The Autostadt triggers as the identification location of a momentum, that which is required of good advertisements (Jung & von Matt, 2003). In doing so, it takes in the danger of a novel logo-centric aberrance, as the brand, for which the logo is emblematic, outshines the product (Klein, 2001). What happens here is the becoming of a metaphor. The BrandLand, with its architecture, becomes a metaphor for the contents, the values of a corporate group, and simultaneously for the manner in which the underlying IT will be deployed in the postmodern.

The implementation of these requirements in the framework of the superior philosophy of a customer-oriented, personal technology seems to be a success. With examples such as the CCC (Customer Care Center) and the IAS (Integrated Autostadt System), the changed awareness with regard to the integration of technology in the personal and organizational relationships allows itself to be realized. It becomes clear that the path to a new, more emotional technology is possible and necessary. The Autostadt takes the first important steps on this long path. Beginning with the installation of networking and dataflow, which are not visible anywhere on the grounds of the park, to the artistic arrangement of the exhibits, which utilize the technical possibilities of imagery, sound and movement without seeming intrusive, and even with the Internet presence, which gives preference to aesthetics over self-presentation, technology gains acceptance and meaning in the service to ideas and content.

THE DIGITAL HEART: INVISIBLE NETWORKS

Embedded in a concept of future-oriented service is where the heart of the Autostadt beats. The “digital heart” is thereby probably only a paradox. Technology no longer presents itself, but rather it puts itself in the service of the people and subordinates to them. Digital technology, originally considered to be cold and sober, should become communicative and people-oriented, only to appear on the surface when necessary. This new “warmth” of the digital world can be transferred from the visitor to the relationship to the Volkswagen Group, and serves as a branding strategy based upon emotions. This is related to the paradigm shift in which technology is expected, regardless of its importance, to adapt to the people and act in the background like a good butler or referee.

The beautifully maintained park should have a convincing effect on the guests, positively touching the senses. Together with the architecture, it subliminally conveys an important message of the corporation; perfection. Thereby, it should convey a constant ability to renew itself. This breaking-out of the limits of style has the goal of higher quality and better service: the guest should feel well taken care of, safe and understood. But in order to ensure that this “emotional work” produces the “right” emotions, the IT systems are trusted with the responsible task of ensuring that the customer is served flawlessly. As the computer, with its algorithmic, binary-coded thinking, simplifies the concept of rationality, IT in the postmodern society can once again combine information with emotion and come across with a humanized approach (Glaser, 1994). “The chip-revolution,” says Glaser, “with its algorithmic, binary-coded thinking, made the concept of rationality radically one-sided”
Emotional Digitalization as Technology of the Postmodern

—with the danger of dehumanization, but yet the postmodern society still combines information and emotion (Glaser, 1994).

With the example of the Autostadt, it is recognizable that steps for a realization of the postmodern are just passable in the area of IT. Whereas the postmodern is concerned comparatively little with reality, the people-conscious deployment of IT shows that the path to a new, more emotional technology is possible and necessary. Endeavours such as the Autostadt take the first steps on this long path. To a far greater extent than in previous decades, today one’s wishes, demands and requirements constitute the parameters of successful process engineering.

The human aspect, particularly in such a service-oriented environment, assumes absolute priority. Thus, the Autostadt is not to be understood as a demonstration of technology; but rather as putting itself in the service of, and subsequently subordinating itself to, the people and the demonstration. In other words, without IT much of the Autostadt would not have been realizable, and thereby cannot be dismissed, but still presents itself in a gentle understatement. Due to this fact, it is explainable why the CTO (Chief Technology Officer) of the Autostadt is not an IT-specialist, but rather pursues psychodynamic and organizations-logical aspects in the exposure of IT in the Autostadt and thus sees his emphases in this area.

**IT AND POSTMODERNITY IN THE AUTOSTADT**

The Autostadt is, as the communication platform of the Volkswagen Group, the actual location of direct dialog between the producer and the consumer, the meeting point for the VW Group and its brands. The values of Volkswagen and the delivery of its automobiles will be “staged” here. The “stylistic device” of the Autostadt is infused with emotions and the urbanistic overall concept is based on the paradigm “structure and event.” Every square meter is staged, and nature and technology meet in a dialog, which is significantly supported by IT.

The IT in the Autostadt constantly produces new, surprising, emotion-awakening effects. For this reason, it is correct to assume that the actual core process of the Autostadt is the “process of generating feelings,” which should then be subsequently visualized. Based on its philosophy, the Autostadt pioneers new and surprising trails in the world of IT. A paradigm shift takes place, which can rightly be referred to as “emotional digitization.”

The concurrence of park, buildings, exhibits, staff and events results in a unity of the main components of emotional stimulants at the Autostadt. The conveyance of perfection is an attribute for which most individuals strive. In this sense, the charisma of the park positively influences the senses and subliminally conveys an important message of the corporate group. The buildings are also in accordance, as they portray a measure of perfection. They are constructed as entirely modern and stylistic. The visitor discovers methods of construction which are seldom, if ever, to be seen in normal surroundings. Beside the perfection, a new effect develops: the progress, orientation to new pathways, and the ability to continually reinvent itself.

With images, metaphors and sequences, the corresponding philosophies of the brands will be conveyed in the individual pavilions, giving further impulses to the above characteristics. In addition, in the events, the impossible will often be made possible with style. Further stylistic devices are also present in the perfection and progress. The Autostadt would like to inform the customer of new trends, acting as the trendsetter for the corresponding impulses.

For example, the POIs (Points of Information) are central components of the information landscape, but in the Autostadt, the POIs are elements of the Lebenswelt. The highly complex network
Emotional Digitalization as Technology of the Postmodern

computer with its front-end software allows itself to be intuitively understood and operated. The guests are not only using systems, but they are experiencing technology. The POIs approach the visitor and they adapt to the user’s size and position. Direct contact with the touchscreen can be seen as a metaphor for human interaction with the invisible network of information technology. As “emotional engineering,” this connection can be understood as a link between the complexities of IT, human gestures and aesthetics. The physically present technologies undergo an emotional exchange with the intellectual techniques to allow the visitor interactive comprehension.

We strive for understatement — that is how we see post-modern IT, which is what makes it possible to “produce” emotions in this context in the first place. Digital technology must be communicative and people-oriented. The complex system structure clearly shows the workload involved in making IT “invisible.” How is it possible to control something that is nearly invisible? Or to state it more clearly, how is the Autostadt itself capable of identifying possible malfunctions at the earliest point in time? Solutions have been found for this. With the bidirectionally media control system “Creston,” the presentations and the technology are systematically monitored. Thus the technology becomes self-referential.

All employees of the Autostadt have access, via PC workstations or central lounges for the Web-based intranet, to all information, e-mail and the Internet. The technology would better be described as “experienced” rather than “used.” This proves that the possibilities of digitization can be used and exploited to prepare the visitor for what they will experience. The invisible, unobtrusive digital elements can help make sure people will not be forgotten, even in a technological environment. Even the demand of philosophy for avant-garde art “within scientific-technical modernization rationality” can only be fulfilled where technology creates space (and scope) for art and aesthetics.

THE FUTURE OF IT

Has postmodernity found its end in the attempt to unify its differences? Have we at least reached a new level of the postmodern, with our humanization-oriented technology? And we continue to ask: where will information technology go from here? We no longer pull levers, we push a virtual button on a touchscreen and in the real world a reaction follows; a lamp goes on, a door closes or a lawn sprinkler starts watering. The same fascinating moments offered by technology in its early mechanical stages are now being provided by virtual technology. Is modernity simply repeating itself on a higher, more abstract level, or is this the beginning of something new?

In the interactive worlds of virtual media, something like a metaphor crystallizes itself out of reality. Due to IT, we experience and learn things that would normally be far out of reach. And what comes along are the new medias demand and encourage a strengthened inclusion of aesthetics in our daily life, and as each is more unobtrusive, its effect becomes stronger. Thus, the success of IT lies in its inclusion in our daily life, which must take place without obstruction and recognition.

The Internet presence of the Autostadt connects a demand of aesthetics to the POIs. The design quality, in which both Web-based systems will communicate informative content and interactive elements, has already been awarded prizes from, among others, New York and Cannes. In an intranet, which is (regarding its public sister) conceptionally and aesthetically behind in nothing, the employees can inform themselves about all-important concerns of the corporation. Standards will also be set here, almost unnoticeable to the public.

The formative characteristic of this aesthetic demand is not the eye-catching distinctive feature, but rather the opposite: the enjoyable appearance, the pleasing control of the view and the (visual) self-explanatory function and order, which lie
behind the contents. These aesthetics in the simulation of contents is effectively “unmodern” in a cultural sense, as a principal departure from rationalization. The people experience IT, they use it subconsciously. IT, as a medium, will be internalized and no longer understood and perceived as a “machine,” a tool, or means to an end. IT will become a medium for emotions and information. The visualized aesthetics of virtual worlds serve in the same way as the aesthetic architecture in the Autostadt.

One could also see the Autostadt as a metaphor, but it was not planned this way. It is a frame of reference for the characteristics and symbols of the postmodern world (not only for the area of IT). Everything – the automobile, as well as the art – will be presented in a different, atypical context. The character systems interlock virtually, and what results is a BrandLand as a liminal space, an area of the threshold, a location of change and the intervention – the barely apprehensible space between the world, which causes a threshold experience of the senses. A location, where we can foretell the future, without being able to see it.

What remains? The IT has taken a significant step: it has become a production factor; it has retreated to the background of business processes, because an IT-networked implementation is the prerequisite for “the social dimension of processes.” The workforce can no longer continue to work autistically. Interdisciplinary qualities such as commercial knowledge and organization talent are vital, at least in strategically decisive positions. The supporting IT does not serve as only the acceleration of postmodern phenomena such as globality and mobility, but it also serves – in a far more readily available manner – as the bridge between the aestheticizing and emotionalizing of our everyday world!

But it will not stop here. It is conceivable that IT in future mechanical engineering will not exist as a purpose in its own right, but will produce tools (systems) for the generation of products, to transfer knowledge. This forces everyday aspects back in the place of revolution, and finally integrates IT permanently into our postmodern world. In spite of all the risks of abuse and autonomy that accompany IT, the possibilities are there.

REFERENCES


Emotional Digitalization as Technology of the Postmodern


ENDNOTES

1 Cp. Jean-Francois Lyotard: Der Widerstreit. München, 1989. pp. 225-226: “Don’t ‘we’ explain ourselves anymore — and let it be with bitterness or rejoicing — the great explanation from the end of the great explanations? Is it sufficient that the thinking of the end of history thinks according, so that it remains modern? Or is post-modernity the business of an old man who searches through the trash can of usefulness for leftovers … and therefore … wins his promise of change!”

2 Peter V. Zima: Moderne–Postmoderne. Gesellschaft, Philosophie, Literatur. Tübingen, Basel 1997. p. 18 – For Zima, “post-modernity” is embossed by the crisis of the modern value system. – The term “Lebenswelt” encompasses the entire spectrum of an individual’s or society’s life, experiences and understanding.

3 Umberto Eco: Nachschrift zum “Namen der Rose.” München, 1986. p. 77: “One could say that every epoch has had its post-modernity, as one had said that every epoch has had its Mannerism (and perhaps, I ask myself, is post-modern the modern name for Mannerism as a meta-historical category).”


Emotional Digitalization as Technology of the Postmodern


Cp. Alfred Schütz/Thomas Luckmann: Strukturen der Lebenswelt. Bd.2. Frankfurt/Main, 1984. p.197: “The symbolic meanings are also – fixed to certain objects bearing significance – memories of experiences in atypical realities, which are brought from other states to the normal state of the everyday life.”


Cp. Homi K. Bhabha (Ed.): Die Verortung der Kultur. Tübingen, 2000. p.1: “Beginnings and ends are arguably the primary myths of the middle ages; but in fin de siècle, we find the moment of change, where time and space meet and complex configurations of difference and identity, of past and future, inside and outside, inclusion and exclusion procreate.”


Ibid. p. 12

Chapter 7.19

The Paleolithic Stone Age Effect?

Gender Differences Performing Specific Computer-Generated Spatial Tasks

Geoffrey S. Hubona
Georgia State University, USA

Gregory W. Shirah
National Aeronautics and Space Administration, USA

ABSTRACT

Most computer applications feature visual user interfaces that assume that all users have equivalent propensities to perceive, interpret, and understand the multidimensional spatial properties and relationships of the objects presented. However, the hunter-gatherer theory (Silverman & Eals, 1992) suggests that there are modern-day differences between the genders in spatial and cognitive abilities that stem from differentiated prehistoric sex roles. If true, there may be discrepancies in how males and females differentially utilize particular spatial visual cues and interface features. We report three experiments in which participants engage in visual spatial tasks using 2D and 3D virtual worlds: (1) matching object shapes; (2) positioning objects; and (3) resizing objects. Female subjects under-perform male subjects in the matching and positioning experiments, but they outperform male subjects in the resizing experiment. Moreover, male subjects make more use of motion cues. Implications for the design of gender-effective user interfaces and virtual environments are considered.

INTRODUCTION

A perennial trend in the evolution of computer technology relates to the ever-increasing power of hardware and the resulting burgeoning possibilities to develop more complex software. These trends have enabled the proliferation of more specialized and powerful computer applications that support users in a wide variety of personal...
and professional tasks. Associated with these trends are multiple challenges: (1) to make the presentation of geometrically increasing amounts of data ever more concise; and (2) to condense, convey, and present larger and larger volumes of useful information using smaller and smaller spaces. To meet these challenges, new and creative approaches to the design of visual user interfaces have emerged, many that present complex, multidimensional data sets and relationships into condensed visual forms and spaces.

Unfortunately, an implicit assumption in the design of commonplace visual user interfaces is that preponderant portions of the existing user population have similar abilities to cognitively perceive, process, interpret, and ultimately understand the intended visual and spatial properties of the objects presented. Yet, it is known that certain measures of spatial cognition are correlated with performance in user interface tasks (Cockburn, 2004). As an example of how individual perceptual differences can affect user interface design, professional Web designers are aware of color blindness patterns in the general population that affect the ability to correctly perceive color-encoded information. As a result, professional designers of high-traffic Internet Web sites avoid these color blindness traps in order to enhance the universal usability of the sites.

Clearly, the assumption of equivalent user capabilities runs the risk of impairing the usability of visual interfaces that ignore broad, existing population anomalies in spatial cognitive and task performance abilities. Through the process of evolutionary natural selection, the hunter-gatherer theory (Silverman & Eals, 1992) ties modern-day, gender-based differences in certain cognitive, spatial abilities back to sharply differentiated sex roles from prehistoric times. Also, it is recognized in behavioral research communities that there are innate differences between the male and female genders related to cognitive spatial abilities (Kimura, 2000; Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995). These gender differences may directly impact the ability to perceive, interpret, and cognitively process spatial properties and spatial relationships of multiple visual objects presented on a computer screen. Thus, there may be fundamental differences between the genders with respect to the ability to use certain visual user interface features, particularly when these features relate to the perception of depth and to the spatial relationships of objects and scenes presented at varying levels of intended depth.

In this article, we review theory and empirical studies relating to (1) gender and human computer interaction and (2) gender differences in innate spatial cognitive and task performance abilities. We then describe three experiments that examine gender-based performance differences in object matching, positioning, and size estimation tasks using two-dimensional (2D) and three-dimensional (3D) virtual worlds. The observed gender performance differences are discussed with respect to applicable theory and with respect to the design of gender-neutral user interfaces and virtual environments.

THEORY AND BACKGROUND

Gender and Human-Computer Interaction

Researchers long have acknowledged the relevance of gender as impacting human computer interaction. Gender has been noted as a broad issue affecting computer skills and computer design issues (Balka, 1996). Gender has been recognized as an important consideration for the design of user interfaces (Leventhal, Teasley, & Stone, 1994) and display techniques (Shneiderman, 1990) and as an issue relevant to achieving universal usability among diverse users of Web-based computer services (Shneiderman, 2000). Gender has been related to the process of decision making, to preferences for investment models, and consequently, as an important consideration.
in the design of financial (Palma-dos-Reis & Zahn, 1999) and organizational decision support systems (Powell & Johnson, 1995) for men and women. It has been shown that there are different perceptions and preferences between men and women with respect to the use and satisfaction with different features of electronic commerce Web sites (Simon, 2001).

Numerous researchers have noted differences between the genders while interacting with computers. For example, it has been shown that boys and girls think differently about computers (Hall & Cooper, 1991; Wilder, Mackie, & Cooper, 1985), and that boys and girls have different motivations for using computers (Inkpen et al., 1994; Upitis & Koch, 1996). Moreover, gender-specific preferences for computer interface features and usage styles also have been documented (Lockheed, 1985). Hinckley, Pausch, Proffitt, and Kassell (1998) reported that females were faster than males performing a two-handed, 3D neurosurgical visualization (manipulation) task, and it has been suggested that females outperform males at certain dexterity tasks (Halpern, 1986; Hinckley, Pausch, Proffitt, Patten, & Kassell, 1997).

However, some studies have found no gender differences while interacting with computers. For example, Inkpen (2001) found no boy-girl differences in children’s interaction styles with point-and-click as compared to drag-and-drop interfaces. In addition, Riemann (1996) found no significant gender impact on the number of reported exploratory learning discoveries using new systems.

One area of HCI that has examined closely the gender performance differences is with respect to the exploration, use, and navigation of virtual reality (VR) and virtual environment (VE) applications. Investigating gestural input techniques for multimodal and virtual environment applications, Wexelblat (1995) reported that gender was not a reliable predictor of gesture frequency for subjects describing movie scenes. Basdogan, Ho, Srinivasan, and Slater (2000) found that female subjects engaged in haptic communication in shared virtual environments (SVEs) reported higher levels of sense of togetherness in performing a collaborative task with an unseen partner than did male subjects. Kauppinen, Kivimaki, Era, and Robinson (1998) argued that gender differences in interacting with others in Collaborative Virtual Environments (CVEs) can be traced to broader, societal-driven, gender-specific, identity distinctions that are also witnessed in natural, non-computer-mediated interactions.

Waller, Hunt, and Knapp (1998) suggested that the transfer of spatial knowledge in virtual environment training is responsible for males outperforming females in computer-generated environments. Similarly, other studies have reported men outperforming women (Astur, Ortiz, & Sutherland, 1998) and making use of different cues than women (Sandstrom, Kaufman, & Huettel, 1998) in navigating virtual worlds. Tan, Robertson, and Czerwinski (2001) reported that men completed 3D virtual environment navigation tasks more quickly than women while using smaller, 15-inch displays, but that this male performance advantage disappeared when using larger, 39-inch displays. Subsequently, Czerwinski, Tan, and Robertson (2002) extended this work with two navigation studies. The first study replicated their findings that a wider field of view combined with a large display reduces gender performance biases. Their second study demonstrated that wider fields of view assist females’ performances in navigating virtual worlds that are densely populated with objects.

**Gender Differences in Spatial Abilities**

Innate gender differences in mental spatial abilities generally are identified as the basis for gender disparities in performing spatial tasks (Kimura, 2000; Linn & Petersen, 1985; Voyer et al., 1995). Certain meta-analytic studies (Linn & Petersen, 1985; Voyer et al., 1995) do indicate a male ad-
The Paleolithic Stone Age Effect?

Vantage on particular cognitive spatial tests, but individual studies’ results are inconsistent in this regard. Further, the different studies often use varying test instruments to measure spatial abilities. Linn and Petersen (1985) categorized the various instruments reported in the literature into three distinct groups: those that measure (1) spatial perception, (2) mental rotation, and (3) spatial visualization. Spatial perception is described as the ability to determine spatial relations despite distracting information. Mental rotation refers to the ability to rotate quickly and accurately two- or three-dimensional figures in imagination. Spatial visualization is the ability to manipulate complex spatial information when several stages are needed to produce the correct solution. These meta-analyses (Linn & Petersen, 1985; Voyer et al., 1995) conclude that men score higher than women on spatial perception and mental rotation cognitive tests, but that neither gender has higher scores on spatial visualization cognitive tests.

The hunter-gatherer theory of the origin of sex-specific spatial attributes (Silverman & Eals, 1992) is one prominent theory that offers an evolutionary perspective on gender differences in modern-day spatial abilities. This theory suggests that men and women have different present-day cognitive skill predispositions that relate to handling differentiated sex role aspects from prehistoric times. Prehistoric females, or gatherers, who could effectively forage for food, and who were successful at keeping track of relationships, activities, objects, locations, and landmarks near their habitats, were superior at acquiring resources for bearing and raising offspring. On the other hand, prehistoric males, or hunters, who could travel better in unfamiliar territory, estimate distance, and navigate with a bird’s-eye view orientation, were, as a consequence, more successful at hunting, competing with other males, finding mates, and, thus, fathering offspring.

The hunter-gatherer theory suggests that these male-female cognitive predispositions persist today through the process of natural evolutionary selection. As evidence supporting this theory, it has been shown that contemporary females outperform men on spatial tasks related to foraging-related activities, such as remembering the location of objects (e.g., landmarks) in their environment (Dabbs, Chang, Strong, & Milun, 1998). Moreover, it has been demonstrated that women outperform men at keeping track of objects and in finding objects that are lost (Eals & Silverman, 1994; Silverman & Eals, 1992). In addition, studies have shown that women remember the locations of previously viewed items better than men (McBurney, Gaulin, Devineni, & Adams, 1997), and that women outperform men remembering the locations of specific objects (James & Kimura, 1997). In contrast, men typically outperform women at spatial tasks manipulating objects in space (Collins & Kimura, 1997; Goldstein, Haldane, & Mitchell, 1990; Kimura, 1983; Kolb & Whishaw, 1990; Linn & Petersen, 1985; Lohman, 1986; Maccoby & Jacklin, 1974). Other studies have demonstrated that men have more adept mental rotation spatial abilities than women (Dabbs et al., 1998; Silverman, Choi, Mackewn, Fisher, Moro, & Olshansky, 2000), purportedly as an evolutionary artifact of the ability to pursue an animal through unfamiliar terrain and then expeditiously find their way home.

**METHOD**

**Hypotheses and Experimental Tasks**

We report three experiments performing spatial tasks using 2D and 3D virtual worlds: object matching, object-positioning, and object resizing. The object-matching experiment was designed largely to tap mental rotation abilities (Linn & Petersen, 1985). The object-positioning and resizing experiments were designed to tap spatial visualization abilities (Linn & Petersen, 1985). Commensurate with the spatial abilities literature, we expect men to outperform women.
in the object-matching task, but there should be no male-female performance differences in the object-positioning/resizing tasks. Consequently, we propose the following two hypotheses:

H1: Male subjects will outperform female subjects matching objects using a mental rotation paradigm.

H2: Male and female subjects will exhibit equivalent performances positioning and resizing objects.

**Object-Matching Experiment.** The object-matching experiment was based on the mental rotation paradigm first developed by Shepard and Metzler (1971). Viewing successive pairs of object images presented from different angles, the task was to judge as accurately and as quickly as possible whether the two images represented identical or different objects. For example, Figure 1 shows a typical object-matching image pair. As quickly as the subject could judge whether the split-screen image pair represented the same or different objects, she or he clicked a corresponding *same* or *different* button on the interface, causing the next image pair trial to be presented. Exactly one-half of the 208 randomly presented image pairs represented identical objects, and the other half showed non-identical objects in the pair.

One-half of all trials were viewed by the men and women subjects in stereo, using 3D Crystal-Eyes™ glasses. The remaining trials were viewed in 2D (monoscopically). The left object image in each trial was always stationary, while the right image was always capable of motion and, specifically, two kinds of motion: (1) in one-half of the trials, subjects could control the motion of the right object image by rotating it in any direction for 360 degrees around the center; and (2) in the remaining trials, the right object always rotated automatically in a fixed and random direction about the center point, rotating at a constant speed of approximately 18 degrees per second. The measured performance variables included *error rate*, the percentage of incorrect matching responses, and *response time*, measured in milliseconds.

Seventeen males and 14 female subjects volunteered to participate in the object-matching experiment. All subjects were employees or contractors of the Goddard Space Flight Center in Greenbelt, Maryland. Subjects with corrected vision wore their eyeglasses underneath the stereoscopic
viewing glasses. All subjects had professional occupations and included engineers, computer programmers, and computer scientists. The mean age of the subjects was 34.97 years with 4.48 mean years of education beyond high school, 17.03 mean years of computer experience, and 12.65 mean years of professional work experience.

Object-Positioning Experiment. The object-positioning task consisted of subjects viewing computer-generated virtual worlds containing three identically sized spherical objects suspended in 3D space (see Figures 2 and 3). As quickly and accurately as possible, subjects were asked to reposition a target object in order to complete a straight line vector configuration defined by three spheres positioned at equal distances from each other. For example, Figure 2 shows a typical initial scene presented at the beginning of a positioning trial. Subjects would fly the object to be repositioned within the virtual world using a (six-degrees-of-freedom) spaceball input device. When satisfied that they had correctly positioned the misplaced object, they pressed a button on the spaceball that recorded their performance data in an output file and caused the next world to appear immediately. Figure 3 shows the correct (solution) placement of the spheres for the initial trial scene presented as Figure 2. Note that correctly performing the positioning task required subjects to locate the displaced object in three dimensions: x (left and right); y (up and down); and z (toward and away from the viewer). One hundred and forty-four unique worlds were presented to each subject in random order.

Thirty volunteer subjects, 14 female and 16 male, participated. All had professional occupations as computer programmers, analysts, and scientists at the Goddard Space Flight Center. The subjects’ mean age was 35.03 years, with 5.53 mean years of education beyond high school, 17.07 mean years of computer experience, and 13.12 mean years of professional work experience.

One-half of all positioning trials was viewed stereoscopically using 3D CrystalEyes™ glasses, while the remaining trials were viewed in 2D (monoscopically) without wearing the glasses. Because subjects wore the glasses to view scenes in 3D, the stereo and mono trials were presented in cohesive blocks of 72 scenes each. The starting order for presenting the stereo and mono blocks was alternated between subjects.

The measured performance variables included distance error magnitude, response time, and rotational distance magnitude. Distance error magnitude was defined as the Euclidean summation of the three directional errors in the x, y, and z dimensions, or $((e_x^2 + e_y^2 + e_z^2) ^ {1/2})$. Thus, this metric measured the exact absolute distance of the repositioned target object from its correct location in three-dimensional space. Response time again was measured in milliseconds. Also, subjects were able to voluntarily rotate the world left or right a total of 45 degrees from the center in either direction, using the left and right arrow keys on the keyboard. The total number of degrees in which the world was rotated in both the left and right directions was captured as a performance rotational distance metric. In addition to positioning distance accuracy and response latency, we were particularly interested in observing mean variances in rotational distance as a function of gender. We introduced the ability for the subjects to rotate the worlds for two predominant reasons: (1) to extract another dependent variable performance measure in addition to standard accuracy and response time measures; and (2) to provide an additional motion-related cue and to see if the male and female subjects would use this cue differentially.

Object-Resizing Experiment. The object-resizing task consisted of subjects viewing virtual worlds containing two differently sized spherical objects suspended in 3D space and displaced at different depths from the viewer (see Figures 4 and 5). Unlike the positioning task, the resizing objects were fixed in position. The task was to adjust the size of a target object in order to correspond with the apparent size of a referent object. Figures 4
The Paleolithic Stone Age Effect?

Figure 2. Initial positioning scene trial

Figure 3. Completed positioning scene trial
Figure 4. Initial resizing scene trial

Figure 5. Completed resizing scene trial
and 5 show a typical set of starting and correctly resized ending virtual worlds. One hundred forty-four unique worlds again were displayed in random order to each subject within alternating blocks, consisting of 72 2D (monoscopic) or 3D (stereoscopic) scenes. Each subject viewed the same 144 worlds (although in a random order); thus, the average target and referent ball sizes across all scenes were equivalent and, therefore, the same for all males and females.

The same 30 subjects from the object-positioning experiment participated in the object-resizing experiment. The measured performance variables included radius error percentage in addition to response time and rotational distance magnitude. Specifically, radius error percentage was defined as the absolute value of the difference between the correct (e.g., referent sphere) radius length and the final resized (e.g., target sphere) radius length divided by the radius length of the referent sphere, or 

\[
\frac{\left| (\text{Correct}_RL - \text{Final}_RL) \right|}{\text{Correct}_RL}
\]

An accuracy measure relative to the size of the referent object was used, because the referent objects randomly varied in size from large to very small. The response time and rotational distance performance metrics were identical to those used in the positioning experiment.

**Assessing Gender Differences in Cognitive Spatial Abilities**

Gender-based differences in subjects’ mental (cognitive) spatial abilities were assessed using the factor-referenced cognitive tests (Ekstrom, French, Harman, & Dermen, 1976) developed by the Office of Naval Research (ONR) and licensed for research use through the Educational Testing Service (ETS). In all three experiments, subjects were administered the cube comparisons and paper-folding cognitive tests. The ETS cube comparisons test assesses mental rotation cognitive ability. In this timed test, subjects were presented with image pairs of wooden cubes, or blocks. Each cube had a different letter, number, or symbol on each of the six faces (i.e., top, bottom, four sides) of the cube. However, in each pair of cubes presented, only three (of the six) sides of the cube were visible. The task was to determine whether the pair represented identical (e.g., the same) or different cubes.

Unlike cube comparisons, which assess mental rotation ability, the ETS paper-folding test assesses spatial visualization ability. In the timed paper-folding test, subjects had to imagine correctly the folding and unfolding of pieces of paper with holes punched through them. According to ETS, cube comparisons require only the mental rotation of the cube configurations, whereas paper folding requires both rotation and visualization, defined as performing serial operations on the configuration. Since both cube comparisons and object-matching task performances are based on the mental rotation process, as described by Shepard and Metzler (1971), cube comparison ability should correspond with object-matching task performances. Similarly, paper-folding skill should correspond to task performances in the object-positioning and resizing experiments, since all are based on spatial visualization ability.

**Results**

**Cognitive Abilities Test Scores.** Tables 1 and 2 reflect the mean scores of the male and female subjects by experiment on the cube comparisons and paper-folding cognitive abilities tests. According to Dr. Ruth Ekstrom (by personal correspondence), the appropriate approach to assess these test results is to consider separately for each test the total number of items answered correctly, the number answered incorrectly, and the number omitted, or unanswered. As indicated in Table 1 (and after checking for equal variances in the male-female scores populations), t-test comparisons of male-female scores on each test indicated that for the object-matching subjects, female subjects answered incorrectly significantly more cube comparison items than males. How-
ever, there were no male-female differences in the number of cube comparison items answered correctly or in the number omitted. Moreover, for the paper-folding test, there were no significant male-female differences in the number of items answered correctly, answered incorrectly, or omitted. By answering significantly fewer cube comparison items incorrectly, there is at least some evidence that the male subjects in the object-matching experiment had an advantage over the females in innate mental rotation cognitive ability. However, the paper-folding test scores indicate no significant male-female differences in innate spatial visualization cognitive abilities.

As indicated in Table 2, t-test comparisons of male-female cube comparison test scores for subjects in the object-positioning and resizing experiments indicate that male subjects answered (1) significantly more items correctly than females and (2) significantly fewer items incorrectly than females. However, there were no significant male-female differences in the number of items omitted in the cube comparisons test. Moreover, there were no significant male-female differences in the number of paper-folding test items answered correctly, answered incorrectly, or omitted. Thus, in the object-positioning and resizing experiments, there again is evidence that males

---

Table 1. Mean cube comparisons and paper folding test scores by gender for subjects in the object-matching experiment

<table>
<thead>
<tr>
<th>Object Matching Subjects</th>
<th>Cube Comparisons Test (Mental Rotation)</th>
<th>Paper Folding Test (Spatial Visualization)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>Answered Correctly</td>
<td>26.65</td>
<td>26.00</td>
</tr>
<tr>
<td>Answered Incorrectly</td>
<td>3.47*</td>
<td>6.07*</td>
</tr>
<tr>
<td>Omitted (Unanswered)</td>
<td>11.29</td>
<td>9.93</td>
</tr>
</tbody>
</table>

* Bolded test scores indicate significantly different (at p < 0.05) male-female scores.

Table 2. Mean cube comparisons and paper-folding test scores by gender for subjects in the object-positioning and resizing experiments

<table>
<thead>
<tr>
<th>Object Positioning/Resizing Subjects</th>
<th>Cube Comparisons Test (Mental Rotation)</th>
<th>Paper Folding Test (Spatial Visualization)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>Answered Correctly</td>
<td>28.25*</td>
<td>21.43*</td>
</tr>
<tr>
<td>Answered Incorrectly</td>
<td>2.38*</td>
<td>5.79*</td>
</tr>
<tr>
<td>Omitted (Unanswered)</td>
<td>11.38</td>
<td>14.64</td>
</tr>
</tbody>
</table>

* Bolded test scores indicate significantly different (at p < 0.05) male-female scores.
compared to females had superior innate mental rotation cognitive abilities. However, similar to the cognitive test data for subjects in the object-matching experiment, there again is no evidence of significant male-female differences in innate spatial visualization cognitive abilities.

Object-Matching Experiment Results. The object-matching data were fitted to a repeated measures multivariate analysis of variance model (MANOVA). There were significant differences in both error rate \((p < 0.0001)\) and response time \((p < 0.0001)\) as a function of gender. Males were

Table 3. Mean object-matching error rates and response times by gender\(^*\) for the viewing and motion conditions

<table>
<thead>
<tr>
<th>Object Matching</th>
<th>Error Rate (%)</th>
<th>Response Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditions:</td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>Stereo Viewing</td>
<td>6.79</td>
<td>10.10</td>
</tr>
<tr>
<td>Mono Viewing</td>
<td>11.99</td>
<td>16.00</td>
</tr>
<tr>
<td>Controlled Motion</td>
<td>7.07</td>
<td>11.95</td>
</tr>
<tr>
<td>Uncontrolled Motion</td>
<td>11.71</td>
<td>14.15</td>
</tr>
</tbody>
</table>

* The minimum significant performance differences (at \(p < 0.05\)) between the genders for the object-matching task are (1) mean error rate: 1.51%; (2) mean response time: 0.43 seconds.

Figure 6. Object-matching error rate by gender by viewing mode
Figure 7. Object-matching error rate by gender by type of motion

Figure 8. Object-matching response time by gender by viewing mode
more accurate at judging whether the objects were identical or different (see Figures 6 and 7). The overall mean male error rate was 9.39%, whereas the overall mean female error rate was 13.05%. Furthermore, the male subjects were faster at making these object comparisons (see Figures 8 and 9). Males exhibited an overall mean response time of 12.53 seconds, whereas females responded in an overall mean time of 13.58 seconds.

To further investigate the gender-based impact of the viewing and motion conditions on object-matching performances, the sample then was split by gender and tested for the effects of viewing mode and type of motion on the object-matching error rates and response times for each gender. The data are presented in tabular form in Table 3. Figures 6, 7, 8, and 9 present these data using line graphs that indicate the minimum significant differences (at $p < 0.05$) between the genders for object-matching error rate and response time (noted in Table 3).

Both male and female subjects were more accurate and faster at matching objects when viewing the object pairs in stereo (see Figures 6 and 8). When the males controlled the motion of the right-hand object image, also called the comparison object (recall that the left image was always stationary), they were more accurate than when they did not control this motion (see Figure 7). For females, there was no significant difference in object-matching accuracy whether they controlled the motion of the comparison object or not (see Figure 7). However, both males and females took longer to judge whether the objects were identical or different when they were controlling this motion (see Figure 9).

**Object-Positioning Experiment Results.** The positioning data were also fitted to a MANOVA model. There were significant differences in distance error magnitude ($p < 0.0001$) and in rotational distance magnitude ($p < 0.0001$) as a function of gender. The male subjects were more
The Paleolithic Stone Age Effect?

accurate in the positioning task (see Figure 10). The overall mean male distance error was 0.573 units, whereas the overall mean female distance error was 0.851 units. Furthermore, the male subjects rotated the positioning scenes to a greater extent than did the females (see Figure 12). The overall mean male rotational distance was 171.99 degrees, whereas the overall mean female rotational...

Table 4. Mean object-positioning distance errors, response times, and rotational distances by gender* for the viewing conditions

<table>
<thead>
<tr>
<th>Viewing Conditions:</th>
<th>Distance Error (Euclidean distance)</th>
<th>Response Time (seconds)</th>
<th>Rotational Distance Magnitude (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
<td>Males</td>
</tr>
<tr>
<td>Stereo Viewing</td>
<td>0.515</td>
<td>0.745</td>
<td>17.71</td>
</tr>
<tr>
<td>Mono Viewing</td>
<td>0.631</td>
<td>0.957</td>
<td>22.49</td>
</tr>
</tbody>
</table>

*The minimum significant performance differences (at p < 0.05) between the genders for the object-positioning task are (1) mean Euclidean distance error: 0.0377 units; (2) mean response time: 0.78 seconds; and (3) mean rotational distance: 9.62 degrees.
The Paleolithic Stone Age Effect?

Figure 11. Object-positioning response time by gender by viewing mode

![Object Positioning Response Time by Viewing Mode](image1)

Figure 12. Object-positioning rotational distance by gender by viewing mode

![Object Positioning Rotational Distance by Viewing Mode](image2)
distance was 150.78 degrees. The difference in positioning response time as a function of gender was not significant \( (p = 0.97) \). Both males and females exhibited a mean positioning response time of 20.1 seconds (see Figure 11).

To further investigate the gender-based impact of the viewing conditions on object-positioning performances, the sample again was split by gender and tested for the effects of mono and stereo viewing on positioning accuracies, response times, and rotational distances for each gender. The data are presented in tabular format in Table 4. Figures 10, 11, and 12 present these data using line graphs that indicate the minimum significant differences \( (at \ p < 0.05) \) between the genders for object-positioning radius error, response time, and rotational distance (noted in Table 4). Both males and females were more accurate and faster at positioning objects in stereo (see Figures 10 and 11). Stereo viewing particularly improved the positioning response time for males more than for females (see Figure 11). Furthermore, both males and females used less rotational distance positioning objects when viewing the objects in stereo, although males exhibited more of this effect than did the females (see Figure 12).

**Object Resizing Experiment Results.** In the resizing task, there were significant differences in radius error percentages \( (p < 0.027) \) and rotational distance magnitudes \( (p < 0.006) \) as a function of gender. The male subjects were less accurate resizing the objects than were the females (see Figure 13). The overall mean male radius error percentage was 15.14\%, whereas the overall mean female radius error percentage was 14.05\%. However, the male subjects rotated the resizing scenes to a greater extent than did the females (see Figure 15). The overall mean male rotational distance was 112.68 degrees, whereas the overall mean female rotational distance was 105.43 degrees. Furthermore, males and females were equally fast at performing the resizing task (see Figure 14). The difference in resizing response time as a function of gender was not significant \( (p = 0.37) \). The mean male response time was 10.68 seconds, and the mean female response time was 10.98 seconds.

The resizing data sample also were split by gender and tested for the effects of mono and stereo viewing on resizing accuracies, response times, and rotational distances for each gender. The data are presented in tabular format in Table 5. Figures 13, 14, and 15 present these data using line graphs that indicate the minimum significant differences \( (at \ p < 0.05) \) between the genders for object resizing radius error, response time, and rotational distance.

---

**Table 5. Mean object resizing radius error percentages, response times, and rotational distances by gender* for the viewing conditions**

<table>
<thead>
<tr>
<th>Viewing Conditions</th>
<th>Object Resizing Performance Measures</th>
<th>Radius Error (%)</th>
<th>Response Time (seconds)</th>
<th>Rotational Distance Magnitude (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Males</td>
<td>Females</td>
<td>Males</td>
</tr>
<tr>
<td>Stereo Viewing</td>
<td></td>
<td>14.61</td>
<td>13.53</td>
<td>9.87</td>
</tr>
<tr>
<td>Mono Viewing</td>
<td></td>
<td>15.66</td>
<td>14.57</td>
<td>11.49</td>
</tr>
</tbody>
</table>

* The minimum significant performance differences \( (at \ p < 0.05) \) between the genders for the object resizing task are (1) mean radius error percentage: 0.96\%; (2) mean response time: 0.64 seconds; and (3) mean rotational distance: 5.20 degrees.
Figure 13. Object resizing radius error by gender by viewing mode

Figure 14. Object resizing response time by gender by viewing mode
The Paleolithic Stone Age Effect?

rotational distance (noted in Table 5). Viewing the worlds in stereo compared to mono had no significant effect on the mean resizing accuracy for either gender (see Figure 13), although it did reduce mean response times for both genders (see Figure 14). Similar to the results from the positioning task, both males and females rotated the scenes less while resizing objects viewed in stereo compared to mono viewing, although the males reduced this stereo-viewed rotational distance more than the females (see Figure 15).

SUMMARY OF FINDINGS

Table 6 summarizes relative task performance differences by gender with each experiment. The spatial abilities literature indicates a robust male performance advantage in mental rotation tasks but no clear advantage for either gender in spatial visualization tasks. Indeed, the ETS factor-referenced cognitive tests administered to our subjects generally corroborate the existing literature. The male subjects answered significantly fewer cube comparisons (mental rotation) test items incorrectly than the females, but there were no differences in male-female paper folding (spatial visualization) test scores.

We hypothesized a male performance advantage in the (mental rotation) object-matching experiment (H1) but no advantage for either gender in the (spatial visualization) object-positioning and resizing experiments (H2). Consistent with the prediction of hypothesis H1, males did exhibit more accurate and faster performances than females in the object-matching experiment. As this task was based on a mental rotation paradigm (Shepard & Metzler, 1971) and coupled with

Figure 15. Object resizing rotational distance by gender by viewing mode

---

Object Resizing Rotational Distance by Viewing Mode

![Graph showing object resizing rotational distance by viewing mode for males and females.](image-url)
spatial literature meta-analyses indicating male advantages in mental rotation tasks and cognitive tests, it is not surprising that the male subjects outperformed the female subjects in the accuracy and speed of matching objects. Moreover, controlling the motion of the right-hand comparison object image significantly improved (i.e., reduced) male object-matching error rates but had no effect on female error rates. Since controlling the motion facilitated the mental rotation task in general, and since males typically outperform females at mental rotation tasks (Dabbs et al., 1998; Silverman et al., 2000), this could explain why males made more effective use of this motion cue.

However, contrary to the prediction of hypothesis H2, in the object-positioning and resizing (spatial visualization) experiments, the relative gender performances were mixed; males were more accurate at positioning objects, whereas females were more accurate at resizing objects. Moreover, the male subjects made more use of the rotational distance feature in both the positioning and resizing tasks. Note that the rotational distance feature, by nature, is a type of motion cue. That is, by rotating the virtual world left or right, the viewer is able to adjust the locations and displacements of the objects in relation to the other objects and in relation to the viewer. The males evidently found this rotating feature to be more useful than females in attempting to position and resize objects in the virtual worlds.

To understand why males were more accurate at positioning objects and why females were more accurate at resizing objects, it is worthwhile to scrutinize the elements of each task. In this regard, object motion was a critical attribute for successfully positioning objects but not for resizing objects. Accurately positioning the target object in a straight line segment required flying the object around the visual space. This essential motion attribute may have contributed to a male performance advantage at positioning objects and is consistent with previous findings that men typically outperform women in tasks manipulating objects in space (Collins & Kimura, 1997; Goldstein et al., 1990; Kimura, 1983; Kolb & Whishaw, 1990; Linn & Petersen, 1985; Lohman, 1986; Maccoby & Jacklin, 1974).

However, the females were more accurate than the males at resizing the target object to match the apparent size of the referent object. The females evidently had a better sense of the relative comparative size of the two objects that were displaced in distance but otherwise fixed in position. It has been demonstrated that females, compared to men, rely more on landmarks for way finding (McGuiness & Sparks, 1983; Miller & Santoni, 1986) and refer to landmarks when

Table 6. Summary of gender-based differences in accuracy, response time, and rotational distance performances for object matching, positioning, and resizing experiments

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Accuracy/Gender</th>
<th>RT/Gender:</th>
<th>Rot. Dist./Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Matching</td>
<td>Males more accurate</td>
<td>Males faster</td>
<td>Not Applicable*</td>
</tr>
<tr>
<td><strong>Object Positioning</strong></td>
<td>Males more accurate</td>
<td>No M/F difference</td>
<td>Males use more</td>
</tr>
<tr>
<td>Object Resizing</td>
<td>Females more accurate</td>
<td>No M/F difference</td>
<td>Males use more</td>
</tr>
</tbody>
</table>

* There was no formal rotational distance dependent variable performance measure in the object-matching experiment. However, when males controlled the motion of the comparison object, they were more accurate (unlike the females) in their object-matching performances.
giving directions (Miller & Santoni, 1986). The literature also indicates that females outperform males in remembering the location of objects that are fixed in space (Dabbs et al., 1998; McBurney et al., 1997). Accurately resizing the target object required the cognitive calibration of the relative apparent sizes of two objects that were displaced in distance from the viewer but otherwise fixed in position. We speculate that the female resizing accuracy advantage could be related to a female propensity to recognize better the locations of landmark objects that are fixed in position (Eals & Silverman, 1994; Silverman & Eals, 1992).

CONCLUSION AND DISCUSSION

In general, the results of these experiments indicate the following:

1. Males outperform females at matching abstract, visual objects using mental rotation in computer-generated virtual worlds.
2. Males make more use of certain motion-related cues in performing visual tasks in computer-generated virtual worlds.
3. Males are more accurate at moving and positioning objects in computer-generated virtual worlds.
4. Females are more accurate at estimating the relative sizes of objects displaced in depth but fixed in position in computer-generated virtual worlds.

How do these findings suggest approaches to developing gender-neutral visual interfaces? The ideal goal in this regard is to move toward the design of user interfaces that will improve usability for both genders and, particularly, to mitigate the postulated female handicap with mental rotation and motion cues. Simply adding a third dimension to an otherwise 2D display does not guarantee improved task performances (Cockburn & McKenzie, 2002). One suggestion is to add meaningful landmarks and to decrease user reliance on mental rotation ability. For example, visual interfaces that allow the stationary user to look right or look left (or up, down, backward, etc.) for familiar landmarks may prove to be an effective, gender-neutral alternative to the use of typical motion cues such as flying around virtual spaces. Our research suggests that the reliance on motion cues, in particular, to extract information can be especially problematic and can lead to a male performance advantage.

An evolving area of visual computing technology to which these findings are relevant is computer-generated virtual environments (VEs). The spatial structure of a VE and the objects that are visible sometimes are used to represent information. According to Vinson (1999), “a VE could contain objects whose spatial properties (e.g., shape, position, size) represent data values on different dimensions. Here, it is necessary for the navigator to quickly develop accurate representations of those spatial properties in order to understand the relationships in the data” (p. 279). These spatial properties (i.e., shape, position, and size) correspond with the three spatial task manipulations analyzed in our studies. To the extent that there are gender biases when cognitively processing and understanding these spatial properties, then these biases likely would extend to understanding the corresponding data relationships in VEs.

Although the tasks analyzed in our studies are not navigation tasks per se, the findings, nevertheless, have implications for the design of navigable VEs, since recognizing object shape and estimating the size and distance of objects are intrinsic activities of successful VE navigation (Vinson, 1999). Navigation is a process of tracking one’s position in an environment in order to arrive at a desired destination. Doing so requires knowledge about positional, velocity, and acceleration attributes, which are derived from location, depth, and motion cues (Cutmore et al., 2000). Location information comes from landmarks and other
fixed, distant features in the environment. The ability to recognize object shapes and to estimate their relative distances (or positions) is necessary in order to acquire basic location information that is a foundation of successful navigation. Moreover, correctly estimating the relative size of a landmark directly relates to understanding its distance and also promotes successful navigation.

Previous studies have indicated male performance advantages when navigating virtual worlds (Astur et al., 1998; Sandstrom et al., 1998; Waller et al., 1998). However, other researchers have noted that the apparent male navigation superiority is mitigated when the virtual worlds are presented to users with wider fields of view and larger display screens (Czerwinski et al., 2002; Tan et al., 2001). Indeed, both genders benefit under these conditions. Tan, Gergle, Scupelli, and Pausch (2003) reported that using larger displays improved both male and female performances in a static, 2D, mental rotation task, but there was no corresponding performance benefit in a reading comprehension task. Subsequently, Tan, Gergle, Scupelli, and Pausch (2004) also demonstrated that users are more effective when performing 3D virtual navigation tasks on larger displays. They noted that when navigating, users continually update their mental sense of position and orientation within the environment (termed spatial updating) using two basic strategies: (1) piloting, or using external landmarks to position themselves within the environment; and (2) path integration, or sensing self-velocity and acceleration to determine their position relative to some starting point. There is evidence that women rely more heavily on the piloting strategy (Czerwinski et al., 2002; Golledge, 1999).

Cockburn (2004) suggested that human spatial capabilities could be leveraged while interacting with computer-generated 3D scenes if the scenes were presented in ways “that better reflect the way we perceive our natural environment” (p. 25). We submit that user interface technology will best support all users to effectively navigate virtual worlds when those worlds are presented to user groups in ways that most closely mirror how users typically and best navigate in natural environments. Flying around over terrain that is visually presented on a limited 2D display screen is not a natural approach for ordinary human navigation. Extending the wider field of view and wider display arguments of Czerwinski et al. (2002), we suggest that the total immersion of users navigating a pedestrian and, apparently, terrestrial virtual world can better support individual human navigation in those worlds. Spatial navigation is improved with an increased sense of presence, or being in the virtual environment, which is a direct outcome of immersion (Tan et al., 2003). For women, these worlds perhaps would be designed best using highly visible and easily recognizable landmarks that are stationary in position and visible within a 360° field of view. That is, women users should be able to look around the virtual landscapes in all directions. Furthermore, the mode of navigating, or movement, should be along the surface of the virtual landscape, especially with alignment and orientation to and from those salient landmarks. For men, cues that further suggest Euclidean distance and direction also should be helpful. Virtual worlds that support both genders’ existing propensities to navigate would be most effective for the majority of users.

**Limitations**

The results of this study should not be interpreted as suggesting that either gender is superior in terms of spatial abilities. Rather, the study suggests (and the reported findings confirm) that there are differences in relative male and female performances on particular spatial processing tasks. Some of these differences are explored in this study. Many others are well-documented in the gender-specific mental abilities literature. In point of fact, both males and females have performance advantages in particular cognitive tasks.
The domain of this study is limited. It focuses on the gender-effective use of abstract visualizations and visual tasks that may relate to particular attributes of visual user interfaces. For example, visual user interface domains that involve the mental rotation of objects or the user-controlled motion, orientation, or resizing of objects, particularly in scenes representing three-dimensional space, are germane. However, domains relating to typical two-dimensional point-and-click activities or to the user interpretation of textual information are not addressed in this study.

In terms of methodology, a limitation of this study is that there was a separate set of subjects in the mental rotation experiment compared to the positioning and resizing experiments. As a result, there was a combination of within- and between-subject comparisons pertinent to the findings and related discussions. However, because these experiments were conducted over a period of time, the attrition of some subjects was unavoidable.

Further, in generalizing these findings, we note that these subjects likely have heightened spatial abilities relative to the general population. They were all working professionals in a premier scientific governmental organization, many holding advanced degrees. Therefore, whether these findings generalize to the population at large is debatable. However, one could argue that any performance effects fostered by the spatial cues would be more pronounced within this group than within the general population. That is to say, if there were no observable effects with these select subjects, then there likely would be no effect within the general population.

Finally, we mention one additional limitation. Egocentric motion tasks, for example, physically walking through a 3D environment, and finding one’s way along a route and/or to and from landmarks have additional perceptual dimensions than do the tasks in our study. However, some elements of egocentric way-finding are embedded in our tasks, such as recognizing the relative shapes, orientations, distances, and sizes of objects in the visual field. Additionally, we did incorporate aspects of egocentric motion skills by enabling the rotation of objects and the rotation of the scenes left and right. However, these tasks do not capture completely all of the perceptual cues and sensory elements inherent in egocentric motion tasks.

**Considerations for Future Research**

There are large segments of the general population that are characterized by differing opportunities, propensities, and abilities to access and make effective use of computer technologies. For example, the very young, the elderly, and the handicapped all have particular impediments to using computer technologies that most people take for granted. As computer applications become increasingly prevalent in everyday life, it is a worthwhile goal to try to mitigate intrinsic barriers and to equalize access and opportunities in order to make effective use of computer technologies.

In terms of gender differences in using visual interfaces, this research is an attempt to identify and highlight broad categories of disparities in male-female abilities to interpret and to manipulate visual abstractions. The intent is to work toward the development of gender-neutral guidelines for effective visual user interfaces and virtual environments. Future research should explore the relative efficacies of different approaches of augmenting computer-generated visualizations so as to promote equal understanding and effective interfaces for different user groups. To this end, investigating approaches that do not rely strictly on abstract visualizations to convey information is warranted. Although it has been demonstrated that visual abstractions effectively can condense and convey large volumes of tabular, multi-dimensional (2D and 3D) data, the combination of visual with, in some cases, textual as well as multi-modal approaches likely will lead to more
effective user interfaces across a larger number of user groups. Furthermore, investigating the design of virtual worlds that most closely mimic how men and women navigate real worlds likely would benefit the largest share of all users.

REFERENCES


The Paleolithic Stone Age Effect?


**ENDNOTES**

1. http://www.webaim.org estimates that as many as 10% of all males and 0.5% of all females are characterized by patterns of color blindness that affect their ability to understand certain shades of color-encoded information.

2. Pooled t-test method was utilized with equal population variances; satterthwaite t-test method was utilized with unequal population variances.
Specifically, they used the Guilford-Zimmer-man spatial orientation task. Please see Tan et al. (2003) for more detailed information.

For a complete discussion of guidelines for designing landmarks to support navigation in VEs, please see Vinson (1999).

Chapter 7.20

Optimality-Theoretic Lexical Mapping Theory: A Case Study of Locative Inversion

One-Soon Her
National Chengchi University, Taiwan

ABSTRACT

Locative inversion verbs seem to share the same argument structure and grammatical function assignment (i.e., <th-OBJ loc-SUBJ>) cross-linguistically. This article discusses the nature of argument-function linking in LFG and demonstrates how the Lexical Mapping Theory (LMT) rendered in Optimality-Theoretic (OT) terms, where argument-function linking is governed by universal violable constraints that consistently favor the unmarked function, accounts for locative inversion straightforwardly. Within this OT-LMT, locative inversion is due to a universal morphosyntactic constraint, and language variation in locative inversion is due to the difference in its relative ranking. This account also offers a potential explanation for the markedness of the locative inversion construction.

INTRODUCTION

The locative inversion construction, as shown in Figure 1, cross-linguistically has similar characteristics in discourse information packaging, which allows the more familiar information to precede the less familiar information (Ackerman & Moore, 2001b; Birner, 1994; Cheng, 1983; Tan, 1991). Between the canonical construction in Figure 1a and the inverted form of 1b, along with the switch of focus from the locative to the theme, is the change in syntactic function assignment. An example from Chinese is given in the figure. The theme role in Figure 1a is assigned the subject function and locative an oblique function; the canonical linking is, thus, <th-SUBJ loc-OBL>. In the inverted Figure 1b, however, the locative is the subject, while the theme now occupies the object position (Her, 1990; Huang, 1993; Huang & Her, 1998; Tan, 1991).
This \textit{<th-OBJ loc-SUBJ>} argument-function “mismatch” was first identified and convincingly argued for in locative inversion verbs in Chichewa (Bresnan, 1994; Bresnan & Kanerva, 1989) and in English (Bresnan, 1989; Tan, 1991). Examples in Figure 2 are from Bresnan and Kanerva (1989).

The subjecthood of the inverted locative phrase \textit{tai-shang} “stage-top” in Figure 1b is evidenced by the fact that it is a bare NP and occupies the usual position for subjects. This is further confirmed by the usual raising test. As shown in Figure 3, \textit{tai-shang} “stage-top” is, indeed, the raised subject, while the “demoted” theme in the post-verbal position, also a bare NP, must be recognized as the object (see Figure 3).

Chinese data thus further confirm Bresnan’s (1994) observation that cross-linguistically, locative inversion verbs share an identical argument
structure <th loc> and the function assignment of the canonical <th-SUBJ loc-OBL> in Figure 1a and the inverted <th-OBJ loc-SUBJ> in Figure 1b. This article aims mainly to account for the syntactic assignment of the argument roles in locative inversion verbs.

In any syntactic theory that aims at characterizing UG, it would be a considerable compromise to simply leave the syntactic assignment of argument roles to lexical idiosyncrasies (Pesetsky, 1995). This article focuses on how the syntactic assignment of argument roles is accounted for universally in the syntactic theory of Lexical Functional Grammar (LFG). This article is organized as follows. The first section discusses how argument-function linking is accounted for by the lexical mapping theory (LMT) in LFG. We will also demonstrate how certain versions of this theory do not account for the locative inversion data from Chinese and English straightforwardly. Furthermore, we will demonstrate how the theory can be improved upon for more consistency and computational efficiency. In the second section we propose a revised LMT formulated as declarative constraints in Optimality-Theoretic (OT) terms. The locative inversion data from Chinese is then accounted for in the third section. The fourth section consists of a discussion on the implications of this study, and the fifth section concludes the article.

The goal of the article is, thus, two-fold: (1) to come up with a universal lexical mapping theory based on violable declarative constraints in OT terms; and (2) to account for Mandarin locative inversion within this comprehensive OT-LMT.

**LEXICAL MAPPING THEORY**

LFG mainly posits three distinct, parallel planes of grammatical description: the argument structure, the functional structure, and the constituent structure (Bresnan, 2001; Dalrymple, 2001; Falk, 2001). The argument structure, or a-structure, consists of the predicate’s thematic and non-thematic argument roles, while the constituent structure, or c-structure, represents the configurational structure, which is the surface structure and allows no syntactic derivation. The functional structure, or f-structure, is the locus of grammatical information, such as grammatical functions (e.g., SUBJ and OBJ), case, person, number, gender, and so forth. The linking of these structures, each with a distinct formal nature, is constrained by correspondence principles. The lexical mapping theory (LMT) is the UG component that constrains the linking between a-structure roles and f-structure functions. The f-structure thus can be viewed as the interface level that links the a-structure and the c-structure. An argument role thus is linked to a grammatical function in the f-structure, which, in turn, is linked to a certain c-structure configuration. The lexical mapping theory (LMT) is the subtheory within LFG, which constrains the syntactic assignment of a-structure roles.

The pioneering work by Levin (1987) started the exploration of more principled accounts to replace the earlier stipulated function-changing rules in LFG. The first comprehensive formulation of LMT was proposed in Bresnan and Kanerva (1989). Since then, even though the essential underpinning assumptions have remained largely stable, the issue of argument-function linking, especially its precise formulation, has yet to be resolved (Butt & King, 2000). A number of different versions of the theory have been proposed (Ackerman, 1992; Ackerman & Moore, 2001a; Alsina, 1996; Bresnan, 1989, 2001; Butt, Dalrymple & Frank, 1997; Her, 1998; Huang, 1993; Zaenen, 1987), among others. A review of these existing versions is clearly outside the scope of this article1. Instead, we will outline mainly the version that seems to be the most widely circulated, found in Chapter 14 of Bresnan (2001), which, in turn, is based largely upon Bresnan and Zaenen (1990)2.
The Theory of A-Structure

Conceptually, LMT consists of two components: the theory of a-structure and the mapping constraints. LFG assumes a universal hierarchy among a-structure roles in terms of their relative prominence in the event denoted by the predicate. This scale descends from the most prominent agent role to the least prominent locative role (Bresnan & Kanerva, 1989, 1992) (see Figure 4).

The most prominent role in an a-structure is called the “logical subject” and is designated \( \Theta \) (pronounced “theta-hat”). In Figure 5, the two-place predicate break requires two argument roles in a-structure, agent (also \( \Theta \)) and theme; the three-place predicate put requires agent (again \( \Theta \)), goal, and theme. Roles in a-structure, by convention, descend in prominence according to the thematic hierarchy.

Grammatical functions (GFs) that can be linked to argument roles are called argument functions. In Figure 6, LFG distinguishes argument functions (shown in bold) from non-argument functions (in italics).

It is important to note that in structure-oriented theories, such as Transformational Grammar (TG) and all its later incarnations, notions such as subject and object are secondary and are derived from structural configurations. In contrast, in relation-oriented theories, such as Relational Grammar (RG) and LFG, these are primary notions in syntax. However, in LFG, argument functions

---

**Figure 4. Thematic hierarchy**

\[
\text{agent} \succ \text{beneficiary} \succ \text{experience/goal} \succ \text{instrument} \succ \\
\text{patient/theme} \succ \text{locative}
\]

**Figure 5.**

\[
\begin{align*}
\text{a. break} & \; < \; x \; \; y \; > \; \; (x = \text{agent}, y = \text{theme}) \\
\text{b. give} & \; < \; x \; \; y \; \; z \; > \; \; (x = \text{agent}, y = \text{goal}, z = \text{theme})
\end{align*}
\]

**Figure 6.**

\[
\text{TOP} \; \text{FOC} \; \text{OBJ} \; \text{OBJ}_4 \; \text{OBL}_4 \; \text{ADJUNCTS}
\]

**Figure 7. Feature decomposition of argument functions**

\[
\begin{array}{c|c|c}
\text{\(\neg\Theta\)} & \text{SUBJ} & \text{OBL}_4 \\
\text{\(\neg\Theta\)} & \text{OBJ} & \text{OBL}_4
\end{array}
\]

\([\sigma] = (\text{an})\text{restricted} \quad [\sigma'] = (\text{an})\text{objective}\)
are further decomposed by two binary features: $[r]$ (whether the function is restricted to having an argument role) and $[o]$ (whether the function is objective) (see Figure 7).

In this system, each argument function is composed of exactly two features and natural classes can be identified, as shown in Figure 8. Furthermore, assuming the minus feature to be the unmarked value, a markedness hierarchy also can be obtained.

Similar to the intrinsic classification of argument roles in Bresnan and Kanerva (1989), Bresnan (2001) assumes that the underlying lexical semantics partially determine the syntactic assignment of different event participants. The universal classification shown in Figure 9 is proposed to capture these predetermined choices of grammatical function assignment.

Cross-language variation in the syntactic assignment of a-structure roles is thus subject to the above universal constraints. The agent role, for example, as a non-patientlike role, is classified $[{-o}]$ by Figure 9c and is thus not associated with OBJ canonically. Patient and theme roles, with the $[{-r}]$ classification, are associated canonically with either SUBJ or OBJ. Under the assumptions in Figure 9, each role in the a-structure is assigned one and only one feature for syntactic function assignment, as morpholexical processes are not allowed to add syntactic features. Language-specific morpholexical operations are allowed, however, to alter the “lexical stock” of an a-structure by adding, suppressing, or binding thematic roles (Bresnan, 2001). Passivization, for example, suppresses $\theta$, the most prominent role, from syntactic assignment (see Figure 10).

In summary, the theory of a-structure renders the argument roles a given predicator requires into an a-structure representation, where roles are listed in a descending order in prominence, and each role is assigned exactly one feature specification for function assignment. The second

---

**Figure 8. Markedness hierarchy of argument functions**

\[
\text{SUBJ}[{r,-o}] > \text{OBJ}[{r,+d}] > \text{OBL}[{-r,-o}] > \text{OBJ}[{r,+d}]
\]

**Figure 9. Semantic classification of a-structure roles for function**

- a. patientlike roles: $\hat{e}$ i $\{r\}$
- b. secondary patientlike roles: $\hat{e}$ i $\{o\}$
- c. other semantic roles: $\hat{e}$ i $\{o\}$

**Figure 10. Passivization**
component in LMT (i.e., the universal set of mapping constraints) then determines exactly which GF each role is assigned to.

**Mapping Principles**

Argument-function linking is subject to certain universal constraints; otherwise, each argument role is freely mapped onto any and all GFs with compatible features. Bresnan (2001) proposes the principles shown in Figure 11.

Two more well-formedness conditions (WFs) are needed in addition to the mapping principles in order to further constrain the non-deterministic argument-function linking: the function-argument biuniqueness constraint and the subject condition:

- Function-Argument Biuniqueness. Each a-structure role must be associated with a unique function, and conversely.
- The Subject Condition. Every predicator must have a subject.

The function-argument biuniqueness constraint ensures a strict one-to-one mapping relation between roles and functions. Computationally, it forces a deterministic assignment to an “unattached” GF between the two GFs with which a role is compatible. The subject condition serves the obvious purpose to ensure that one role in a-structure must be mapped to SUBJ. This condition also forces a deterministic choice when a role’s syntactic assignment is compatible with SUBJ and some other function and when all other

---

**Figure 11. Mapping principles (MPs)**

a. Subject roles:
   
   (i) $s_{x_{a_1}}$ is mapped onto SUBJ when initial in the a-structure; Otherwise,
   
   (ii) $s_{x_{a_1}}$ is mapped onto SUBJ.

b. Other roles are mapped onto the lowest compatible function in the Markedness Hierarchy.

---

**Figure 12.**
roles in the a-structure, if any, are incompatible with SUBJ.

We demonstrate how three different types of verbs receive correct argument-function linking in the LMT just described. An unaccusative verb is given in Figure 12, while an unergative verb is illustrated in Figure 13. A typical transitive verb is given in Figure 14.

**Improvement to the Conventional LMT**

There are several areas at the theoretical level upon which the conventional LMT may be improved. First, the uniform underspecification of each role with exactly one syntactic feature can be relaxed to allow the formalism to be more expressive, yet without compromising its formal power. This classification scheme also may be too rigid in that it does not allow the possibility of agentive objects, which have been observed in several languages (Bresnan, 2001; Dalrymple, 2001). A desirable improvement to the theory is to allow such linking possibilities and, at the same time, be able to express the marked nature of such a linking as agent-OBJ.

As for the mapping principles, two disjunctions are observed. The first one is in the mapping
principles of subject roles: a disjunction exists between $\Theta_{[\alpha]}$ and $\theta_{[\theta]}$, each a stipulation for linking to SUBJ. In a more general theory of UG, it would be desirable not to include such function-specific linking conditions. Notice also the specification that $\Theta_{[\alpha]}$ be the initial role in the a-structure. This principle thus must refer explicitly to the ordering in the a-structure. The second disjunction is found between subject roles and non-subject roles. For the former, a qualified role is mapped to SUBJ (i.e., the most prominent GF). However, on the contrary, non-subject roles must be linked to the least prominent compatible GF. A consistent principle for all roles would make a simpler and more general theory.

Finally, note that the Subject Condition in LFG states explicitly that every clause must have a subject. Similar constraints are also necessary in other syntactic frameworks; for example, the same is accomplished by the Extended Projection Principle (EPP) in Transformational Grammar and the Final-1 Law in Relational Grammar. However, as it often has been noted, such an inflexible stipulation may not be empirically accurate. As cited in Ackerman and Moore (2001a), clauses may truly be without a subject (Babby, 1989; McCloskey, 2001). Bresnan (2001) thus hinted that this condition should perhaps be stipulated as a parameter. Again, ideally, a UG theory should be able to account for such subjectless clauses and their marked nature at the same time.

Assuming that the conventional LMT takes the same position advocated first in Alsina and Mchombo (1993) and does not allow morphologi-
Optimality-Theoretic Lexical Mapping Theory

Optimality-Theoretic Lexical Mapping Theory

As shown in Figure 16, the argument-function linking of <th-OBJ loc-SUBJ> in the locative inversion construction cannot be obtained, even though the canonical linking of <th-SUBJ loc-OBL> is accounted for in Figure 16. Therefore, it would make sense empirically to allow morphological processes in the theory to alter syntactic assignments by adding features, as proposed in Zaenen (1987), Ackerman (1992), Markantonatou (1995), and Her (1998, 2003). The default locative classification employed by Bresnan and Kanerva (1989) and Bresnan (1989), which assigns loc_[r] when th is focused, likewise can be viewed as such a feature-adding morphological operation. In the fourth section, we also will discuss the advantages of feature-adding morphological operations from the standpoint of expressivity and formal power.

In the second section, we will propose an LMT in OT terms, thus an OT-LMT, that attempts to incorporate the desirable improvements suggested here.

**AN OPTIMALITY-THEORETIC LMT**

Optimality Theory has exerted great influence over the field of phonology; however, its application in syntactic theory is still in its infancy. Recently, there have been some explorations within the OT-LFG framework (aka Optimal Syntax [Bresnan, 2000]). From the OT point of view, OT-LFG can be seen as OT with a universal LFG as GEN. From the point of view of LFG, a constraint-based grammatical framework, generalizations are interpreted in OT terms with (violable) constraints ranked in relation to one another (Sells, 2001). A number of studies have been carried out within this general framework (Mikkelsen, 2003; Sells, 2001). There also have been efforts to render argument-function linking in OT terms (Butt et al., 1997; Lødrup, 1999).

**An OT-LFG Overview**

Bresnan (2000) depicts the basic structure of OT-LFG, or Optimal Syntax, where LFG’s correspondence theory of parallel structures serves as a model for GEN. The standard OT-LFG assumes

![Figure 17.](image)
input to be “a (possibly underspecified) feature structure representing some given morphosyntactic content independent of its form of expressions” (Bresnan, 2000). An example is given in Figure 17, which assumes I saw her as its optimal form of expression. Note that in the input structure, \(<x, y>\) is the a-structure of see and GF\(_1\) and GF\(_2\) are unspecified grammatical functions that argument roles with which \(x\) and \(y\) are associated.

The candidate set comprises pairs of f-structure and corresponding c-structure (and perhaps other corresponding planes of information) generated by the LFG grammar (Bresnan, 2000; Kuhn, 2001). For ease of presentation, I am simplifying the matter by taking the input to be an a-structure \(<x, y>\), and a set of \(<x\)-GF\(_1, y\)-GF\(_2>\) pairs as candidates in OT-LMT, which is a module within OT-LFG that constrains argument-function linking specifically. The candidates are evaluated by a universal set of lexical mapping constraints. The output is taken to be the most harmonic, or optimal, candidate pair; namely, the one with the least (serious) violations (Kuhn, 2001).

### A Comprehensive OT-LMT

The OT-LMT proposed here modifies and expands the LMT component in Bresnan (2001) and is based specifically on the particular formulation of LMT in Her (1997), Huang and Her (1998), and Her (2003), where syntactic feature assignments are simplified, and the multiple mapping principles and well-formedness conditions in the conventional LMT are all unified into a single consistent mapping principle. Here, I will take this further and reinterpret the entire simplified LMT as a set of Optimality-Theoretic constraints and thus offer a comprehensive OT-LMT.

Crucial to the theory are two prominence scales discussed earlier: a universal thematic hierarchy and a markedness hierarchy of grammatical functions (GFs) (see Figures 18 and 19).

Mapping constraints are classified into three categories: well-formedness constraints on argument roles, well-formedness constraints on argument functions, and constraints on linking. Note that we are ignoring athematic arguments in this article. “R” is thus a thematic role in a-structure, and “F” is a corresponding grammatical function. We first examine the well-formedness constraints on the representation of argument roles (see Figure 20).

UniqRol ensures the uniqueness of each and every role in the a-structure and thus rules out a-structures like \(<ag ag th>\) and \(<th loc loc>\). DescendRol further formalizes the a-structure representation, where argument roles descend in prominence. For example, given the locative verb sit and its two roles, theme and locative, in a-structure, \(<th loc>\) is the only well-formed representation; \(<loc th>\) is ill-formed. Two corresponding constraints are proposed for argument functions (see Figure 21).

UniqFun ensures the uniqueness of each and every function in the a-to-f mapping; thus, both of the following are ill-formed: \(<\theta_a\)-SUBJ \(\theta_b\)-SUBJ>, and \(<\theta_a\)-OBJ \(\theta_b\)-OBJ>. DescendFun penalizes a candidate with a violation of the descending order in prominence. For example, because SUBJ outranks OBJ, \(<\theta_a\)-SUBJ \(\theta_b\)-OBJ> has 0 violation and is favored over the inverted \(<\theta_a\)-OBJ \(\theta_b\)-SUBJ>, which incurs one violation. Thus, given \(n\) GFs in a candidate form, there are at most \(n - 1\) violations as there are \(n - 1\) consecutive pairs (Kuhn, 2001). Inversion is still possible, given that all OT constraints are violable in order to satisfy higher-ranked constraints, including language-specific morphosyntactic operations. (We will discuss the possibility of a language-specific component in the next section.)

Next, we move on to the general constraints on the linking between roles and functions.

The two constraints in Figure 22, LinkRol and LinkFun, ensure that each expressed role is linked to a GF and that each GF is linked to a role. A role that is not linked to an argument function causes incompleteness, while an argument function that is not linked to an argument role in a-structure...
causes incoherence. Notice that there is no need to specify a constraint just to ensure that a role is linked to a GF with compatible features. This is accomplished automatically by the universal constraints on the morphosyntactic properties of argument roles shown in Figure 23. LinkPtTh reflects the unaccusative hypothesis that cross-linguistically the primary patient/theme

\[ \text{arg} > \text{ben} > \text{gdeSp} > \text{inst} > \text{path} > \text{loc} \]

**Figure 18.**

**Figure 19.**

\[ \text{SUBJ}^{[\gamma] > \text{OBJ}^{[\gamma + \gamma]} \text{OBJ}^{[\gamma - \gamma]} > \text{OBJ}^{[\gamma + \gamma]} \} \]

**Figure 20. Well-formedness constraints on argument roles**

- a. **UniqRole**\( (R_a, R_d) \): Given \( \langle \ldots, R_a, F_a, R_d, F_d \rangle \), \( R_a \) is unique in \( R_d \)
- b. **DescendRole**\( (R_a, R_d) \): Given \( \langle \ldots, R_a, F_a, R_d, F_d \rangle \), \( R_a \) is prominent in \( R_d \)

**Figure 21. Well-formedness constraints on grammatical functions**

- a. **UniqFun**\( (F_a, F_d) \): Given \( \langle \ldots, F_a, F_a, F_d \rangle \), \( F_a \) is unique in \( F_d \)
- b. **DescendFun**\( (F_a, F_d) \): Given \( \langle \ldots, F_a, F_a, F_d \rangle \), \( F_a \) is prominent in \( F_d \)

**Figure 22. General constraints on argument-function linking**

- a. **LinkRole**\( (R, F) \): Given \( \langle \ldots, R \rangle \), \( R \) is linked to an \( F \) such that \( \langle \ldots, F \rangle \).
- b. **LinkFun**\( (F, R) \): Given \( \langle \ldots, F \rangle \), \( F \) is linked to an \( R \) such that \( \langle \ldots, F \rangle \).

**Figure 23. Specific constraints on argument-function linking**

- a. **LinkPrTh**\( (R, F) \): Given \( \langle \ldots, F \rangle \), where \( R = \text{path} \), \( F \) is \( [\gamma] \)
- b. **LinkRolRes**\( (R, F) \): Given \( \langle \ldots, F \rangle \), where \( R \) is \( [\gamma] \) in \( F \)
- c. **LinkUnobj**\( (R, F) \): Given \( \langle \ldots, F \rangle \), \( F \) is \( [\gamma] \)
- d. **LinkUnres**\( (R, F) \): Given \( \langle \ldots, F \rangle \), \( F \) is \( [\gamma] \)
Optimality-Theoretic Lexical Mapping Theory

is encoded as an unrestricted \([-r]\) GF (i.e., SUBJ or OBJ) (Bresnan & Kanerva, 1989; Bresnan & Zaenen, 1990; Zaenen, 1993). LinkRolRes captures the generalization that a non-patient/theme internal argument prefers the syntactic assignment of a thematically restricted function. Finally, LinkUnobj and LinkUnres consistently favor the assignment of a role to the most unmarked function, SUBJ, \([-r-o]\). Each function may have zero to two violations. These two constraints together are more general and insightful than the previous Subject Condition, which simply stipulates that every clause should have a subject.

Note that LinkRolRes does not apply to agent, the external argument. Being the highest-ranked role, it is linked to SUBJ due to LinkUnobj and LinkUnres. This thus accounts for the fact that, for the majority of the world’s languages, agent cannot be realized as an object. However, given the violable nature of these constraints and their variable ranking, the possibility of agent-OBJ does exist as a marked morphosyntactic option. This reflects the insight of Falk (1989) cited in Lødrup (2000) that in Norwegian, “what has been called external theta roles are in fact structurally unspecified theta roles” (p. 173).

-I will follow the standard view in OT and assume that these constraints are universal, but their ranking may be language-specific. For Chinese, I propose the ranking shown in Figure 24.

**An Illustration of OT-LMT**

We will now look at the lexical mapping of three different verbs in their canonical active construction as examples: “melt<th>,” “laugh<ag>,” and “sell<ag th>.” To save time and space, the (many) candidates that violate any of the five highest-ranked well-formedness constraints will be excluded and we only will be concerned with the lower five. Following standard OT notation, in Figure 25, a violation is marked with “*”; a “fatal” violation causing a candidate to lose in evaluation is highlighted with “!” . The shaded area covers the constraints that are no longer relevant in the evaluation of a particular candidate, and finally, the \(\nabla\) sign indicates the optimal selection.

The candidate \(C1, <th-SUBJ>\), with no violation, is clearly the optimal selection, where SUBJ is an unrestricted \([-r]\) function allowed by LinkPtTh, and also the unmarked \([-r-o]\) function preferred by LinkUnobj and LinkUnres.

Next, we turn to the a-structure of an unergative verb “laugh<ag>” (see Figure 26). Here, the only relevant constraints are LinkUnobj and LinkUnres, which again select SUBJ, the unmarked function.

Again, the candidate with no violation (\(CI\)) is the optimal selection.

The final example, *sell*, is a transitive verb with an agent role and a theme role. Again, Figure 27

---

**Figure 24. OT ranking of lexical mapping constraints (Chinese)**

```
UniqRol|DescendRol|UniqFun|LinkRol|LinkFun

LinkUniobj|LinkUnres
```

---

2510
### Figure 25. Input a-structure: ‘melt <th>’

<table>
<thead>
<tr>
<th>C</th>
<th>candid</th>
<th>LinkPrTh</th>
<th>LinkRolRes</th>
<th>DescendFun</th>
<th>LinkUnobj</th>
<th>LinkUnres</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>&lt;th-SUBJ&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>&lt;th-OBJ&gt;</td>
<td></td>
<td></td>
<td>*</td>
<td>!</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>&lt;th-OBL&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>&lt;th-OBJ&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Figure 26. Input a-structure: ‘laugh <ag>’

<table>
<thead>
<tr>
<th>Candidate</th>
<th>L.inkPrTh</th>
<th>LinkRolRes</th>
<th>DescendFun</th>
<th>LinkUnobj</th>
<th>L.inkUnres</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>&lt;ag-SUBJ&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>&lt;ag-OBJ&gt;</td>
<td>*</td>
<td></td>
<td>!</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>&lt;ag-OBL&gt;</td>
<td>*!</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>&lt;ag-OBJ&gt;</td>
<td>*</td>
<td></td>
<td>!</td>
<td>*</td>
</tr>
</tbody>
</table>

### Figure 27. Input a-structure: ‘sell <ag th>’

<table>
<thead>
<tr>
<th>Candidate</th>
<th>LinkPrTh</th>
<th>LinkRolRes</th>
<th>DescendFun</th>
<th>LinkUnobj</th>
<th>LinkUnres</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>&lt;ag-SUBJ th-OBJ&gt;</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>&lt;ag-SUBJ th-OBL&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>&lt;ag-SUBJ th-OBJ&gt;</td>
<td>*!</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>&lt;ag-OBJ th-SUBJ&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>&lt;ag-OBJ th-OBL&gt;</td>
<td>*!</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>C6</td>
<td>&lt;ag-OBJ th-OBJ&gt;</td>
<td>*!</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>C7</td>
<td>&lt;ag-OBL th-SUBJ&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C8</td>
<td>&lt;ag-OBL th-OBJ&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C9</td>
<td>&lt;ag-OBJ th-OBJ&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C10</td>
<td>&lt;ag-OBJ th-SUBJ&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C11</td>
<td>&lt;ag-OBJ th-OBJ&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C12</td>
<td>&lt;ag-OBJ th-OBL&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Optimality-Theoretic Lexical Mapping Theory

excludes candidates that violate any of the five highest-ranked constraints.

Among the candidates, \( C1, <\text{ag-SUBJ th-OBJ}> \), is the optimal selection, even though it does violate one of the two lowest-ranked constraints (i.e., LinkUnobj) due to the linking of theme to OBJ, a function with the marked feature \([+\sigma]\). All other candidates, however, violate at least one higher-ranked constraint. Note that a candidate a-structure where both roles are linked to the unmarked function, thus \(<\text{ag-SUBJ th-SUBJ}>\), violates the highest ranked UniqFun and, therefore, is not included in the figure.

AN OT-LMT ACCOUNT OF LOCATIVE INVERSION IN CHINESE

We first apply the OT-LMT to the canonical a-structure of the locative verb. The theory correctly predicts the following optimal argument-function linking: \(<\text{th-SUBJ loc-OBL}_u>\), shown in Figure 28.

With \( C2, <\text{th-SUBJ loc-OBL}_u> \), as the optimal selection, this constraint ranking obviously does not account for locative inversion, which is represented by candidate \( C4, <\text{th-OBJ loc-SUBJ}> \). An additional constraint is needed.

The constraint shown in Figure 29 draws on the insight found in the default rule for focused theme posited by Bresnan and Kanerva (1989) and also faithfully reflects Bresnan’s (1994) observation on the universals of locative inversion verbs. Given the fact that the complement of the predicador usually carries the discourse function of marking the less familiar information and that the subject is the default grammatical function for topic or more familiar information, the locative inversion operation forces the locative to map onto SUBJ such that the focused theme can surface

---

**Figure 28. Input a-structure: 'sit <th loc>'**

<table>
<thead>
<tr>
<th>Candidate</th>
<th>LinkPrTh</th>
<th>LinkRolRes</th>
<th>DescendFun</th>
<th>LinkUnobj</th>
<th>LinkUnres</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C1 )</td>
<td>&lt;th-SUBJ loc-OBJ&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>( C2 )</td>
<td>&lt;th-SUBJ loc-OBL( u )&gt;</td>
<td>*!</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>( C3 )</td>
<td>&lt;th-SUBJ loc-OBJ( u )&gt;</td>
<td></td>
<td>*!</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>( C4 )</td>
<td>&lt;th-OBJ loc-SUBJ&gt;</td>
<td>*!</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>( C5 )</td>
<td>&lt;th-OBJ loc-OBL( u )&gt;</td>
<td>*!</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>( C6 )</td>
<td>&lt;th-OBJ loc-OBL( u )&gt;</td>
<td></td>
<td></td>
<td>*! *!</td>
<td>*</td>
</tr>
<tr>
<td>( C7 )</td>
<td>&lt;th-OBL( u ) loc-SUBJ&gt;</td>
<td>*!</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>( C8 )</td>
<td>&lt;th-OBL( u ) loc-OBJ&gt;</td>
<td>*!</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>( C9 )</td>
<td>&lt;th-OBL( u ) loc-OBL( u )&gt;</td>
<td>*!</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>( C10 )</td>
<td>&lt;th-OBJ( u ) loc-SUBJ&gt;</td>
<td>*!</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>( C11 )</td>
<td>&lt;th-OBJ( u ) loc-OBJ&gt;</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>( C12 )</td>
<td>&lt;th-OBJ( u ) loc-OBL( u )&gt;</td>
<td>*!</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
as a complement of the locative verb. With this constraint in place, we now have an important decision to make; that is, whether to posit LinLocInv as a language-specific constraint. Recall that locative inversion is found in many languages and locative inversion verbs share an identical a-structure and function assignment. However, locative inversion certainly does not occur in all languages. In non-configurational languages with extensive case-marking for grammatical relations (e.g., Korean and Japanese), locative inversion may not be found (Huang & Her, 1998). Figure 30 depicts a Japanese example.

Notice that even though the locative phrase indeed may invert positions with the subject and thus affect the focus in Figure 30b, its grammatical functions remain the same. In other words, locative inversion does not affect argument-function linking. Recall the standard OT view that constraints are universal and that only their ranking is subject to variation. Therefore, if we follow the standard OT view and posit LinkLocInv as a universal constraint, languages such as Japanese also must be accounted for, but only with a different ranking of the same constraints. This is the path we will explore. Figures 31 and 32 show the revised ranking we propose for Chinese. Notice that LinkLocInv outranks LinkRolRes and is outranked by LinkPtTh. Again, we continue to ignore the five highest-ranked well-formedness constraints.

We also need to point out that LinkLocInv is irrelevant in the selection of the canonical in Figure 28, because there, the theme is not focused. Now, to account for the data from languages like Japanese, where the focused theme does not result

---

**Figure 29.**

\[
\text{LinLocInv}(R, F): \text{Given a-structure } \langle R_a - F_a, R_b - F_b \rangle, \text{ where } R_a = \text{th[fol]} \text{ and } R_b = \text{loc, } F_b \text{ is } [+ - o].
\]

---

**Figure 30.**

a. Henkoputa ga yama no ue ni orimasita.
   helicopter NOM mountain POSS top LOC land
   ‘A helicopter landed on top of the mountain.’

b. Yama no ue ni henkoputa ga orimasita.
   mountain POSS top LOC helicopter NOM land
   ‘On top of the mountain landed a helicopter.

c. *Yama no ue ga henkoputa o orimasita.
   mountain POSS top NOM helicopter ACC land
   ‘On top of the mountain landed a helicopter.’

---

2513
in mismatches of the function assignment of argument roles, we posit the ranking in Figures 33 and 34. Notice here that LinkLocInv is outranked by all other constraints.

**DISCUSSION**

This section discusses three issues in further detail. The first issue relates to the nature and the
scope of the OT-LMT proposed in the article. The second issue concerns the potential advantages that the OT-LMT may have over the conventional LMT. Finally, we explore some of the directions for further research concerning the OT-LMT.

Morphosyntactic vs. Morpholexical Processes

Given the often idiosyncratic nature of language-specific lexical information, it is not yet clear how
the technical integration of the lexicon should be envisaged in OT syntax, in general (Kuhn, 2001). This article clearly does not address this larger issue. In order to have an insightful lexical mapping theory in OT syntax, we first must be explicit about its nature and scope. The OT-LMT envisioned here is part of a universal OT-LFG theory that constrains argument-function linking. In other words, it constrains the syntactic function assignment of argument roles required by a predicator. Thus, this OT-LMT, as it is currently formulated, has nothing to say about morpholexical processes that alter the “lexical stock” in a-structure (Bresnan, 2001; Bresnan & Kanerva, 1989). Therefore, it is purely morphosyntactic in nature and scope. Crucially, Ackerman (1992) differentiates and characterizes morpholexical and morphosyntactic operations as follows:

**Morpholexical (Operations),** affect the lexical semantics of predicates by altering the semantic properties associated with predicates...**Morphosyntactic (Operations),** assign features supplemental to those supplied by IC assignment\(^{10}\): these operations can affect the final GF assignments to arguments but cannot affect the lexical semantics. (p. 56)

Morpholexical operations thus are word-formation processes that produce predicates with an altered inventory of argument roles, or a-structures, which serve as input to OT-LMT. Morphosyntactic operations, however, are within the proper domain of LMT. Assuming that only morpholexical operations may be language-specific, the OT-LMT proposed thus universally governs how argument roles are mapped to GFs, with constraints that may vary from language to language only in terms of ranking. Thus, as Huang and Her (1998) have argued, given the nature of syntactic assignment of argument roles in the theory, it, in fact, makes the theory more coherent by allowing syntactic feature assignment in morphosyntactic operations. This is precisely how we treated locative inversion. Similar proposals that allow morphological processes to affect syntactic assignments by adding features are found as early as Zaenen (1987) and Her (1990) and later in Ackerman (1992), Markantonatou (1995), Her (2003), among others.

Allowing feature-adding morphosyntactic operations, in fact, also offers a computational advantage. Morpholexical operations constitute a much more powerful formal device computationally in that they are not subject to the general monotonicity condition that information only can be added but cannot be deleted or changed (Bresnan, 1990; Falk, 2001).\(^{11}\) Monotonic morphosyntactic operations with the feature-adding capacity enable a formalism that is more consistent and also more expressive, without any increase in its formal power. Empirically, such operations also have been adopted to account for syntactic variations in several languages; for example, Greek (Markantonatou, 1995), Chinese (Huang, 1995; Her, 1999), and English (Zaenen, 1987).

We will now illustrate this view of the OT-LMT with two more constructions from Chinese that are related to locative verbs. The first one is a passivized locative construction. Three-place transitive predicates like xie “write,” with the argument structure \(<ag th loc>\), do not allow inversion in spite of the locative role it requires. However, there may be locative inversion if the agent role is suppressed. This is observed in Chinese (Huang & Her, 1998) and other languages (Bresnan, 1989; Bresnan & Kanerva, 1989). The examples in Figures 35 and 36 are from Chinese and English.

Recall that passivization, repeated in Figure 36, suppresses the logical subject. In effect, it gives rise to an argument structure \(<ag th loc>\), precisely that of a locative inversion verb. Locative inversion, therefore, is allowed, as in Figure 35c. Passivization thus falls outside of the realm of LMT and is regarded as a language-specific operation. One indication of its language-dependence is in the indirect expression of the suppressed
Figure 35.

| a. Amei xie le yi ge zi zai qiang-shang. | Amei write ASP a CL character at wall-top  
| ‘Amei wrote a Chinese character on the wall.’ |
| b. *Qiang-shang xie le yi ge zi Amei. | wall-top write ASP a CL character Amei  
| *‘On the wall was written a Chinese character (by) Amei.’ |
| c. Qiang-shang xie le yi ge zi. | wall-top write ASP a CL character  
| ‘On the wall was written a Chinese character.’ |

Figure 36. Passivization

```
<...>
/
```

Figure 37.

| a. Amei zuo zai yizi-shang. | Amei sit at chair-top  
| ‘Amei sits on the chair.’ |
| b. Amei shui zai diban-shang. | Amei sleep at floor-top  
| ‘Amei sleeps on the floor.’ |

Figure 38.

| a. Amei zuo yizi-shang. | Amei sit chair-top  
| ‘Amei sits *(on) the chair.’ |
| b. Amei shui diban-shang. | Amei sleep floor-top  
| ‘Amei sleeps *(on) the floor.’ |
agent role as an adjunct; for example, the English *by*-expression (Bresnan, 1994). Chinese, however, does not allow such indirect expressions.

The second construction we will examine is the transitivized locative verb. It has been noted that, in Chinese, a two-place locative verb with an argument structure <th loc> in fact allows its locative phrase to be a PP or an NP. The locative phrase thus may be alternatively mapped onto OBL or OBJ (Huang & Her, 1998). This, however, is not allowed in English, as shown in Figures 37 and 38.

Huang and Her (1998) treat this function change as a morphosyntactic variation of the same argument structure; thus, <th-SUBJ loc-OBL> in Figure 37 and <th-SUBJ loc-OBJ> in Figure 38. However, further evidence indicates that this view may be incorrect and that locative transitivization involves a morpholexical change, instead. In other words, without the preposition *zai*, the argument structure is, in fact, no longer <th loc>. Note that the presence of the locative preposition *zai* requires a place noun as its complement in Figure 39. In Chinese, certain nouns are place nouns inherently, such as *xuexiao* “school,” *zheli* “here,” and *gongyuan* “park,” and thus can be the complement of preposition *zai* directly. Non-place nouns, however, must form a constituent with a locative affix such as –*shang* and –*xia*, or a locative noun such as *shangmian* and *xiangmian*; or there

**Figure 39.**

| a. Amei zuo zai yizi-*shang*. | Amei sit at chair-top  
| "Amei sits on the chair." |
| b. Amei shui diiban-*shang*. | Amei sleep floor-top  
| "Amei sleeps on the floor." |

**Figure 40.**

| a. Amei zuo yizi. | Amei sit chair  
| "Amei sits *(on) the chair." |
| b. Amei shui diiban. | Amei sleep floor  
| "Amei sleeps *(on) the floor." |

**Figure 41. Locative transitivization**
must be a place noun as the complement of locative preposition zai. Notice in Figure 40 that the object required by the transitivized verb zuo “sit” and shui “sleep” is free of this restriction.

Therefore, it is clear that the objects in Figure 40 do not denote the location where the theme that undergoes the movement ends up; rather, they are the entities that receive the action denoted by the verbs. To account for this construction, I propose a morpholexical operation (see Figure 41).

Two more syntactic tests, shown in Figures 42 and 43, confirm that this argument structure is now <ag th>: passivization and resultative compounding.

In the passive construction, the suppression of the agent results in the theme role’s “promotion” to SUBJ, as seen in Figure 42. In Figure 43, the single composite role, formed by the binding of the theme role of the action verb and the theme of the result state verb, maps to SUBJ (Her, 2004)13.

Based on the prevailing evidence, locative transitivization should be treated as a morpholexical operation that alters the lexical stock of an argument structure, and, as such, it is again outside of the realm of the OT-LMT proposed here.

According to the previous discussion, it is now possible to indicate exactly how the OT-LMT system is envisaged as the module in LFG that links the lexical semantic structure and the syntactic structure of a predicator (Bresnan & Kanerva, 1989; Bresnan & Zaenen, 1990). The particular conceptualization of the a-structure assumed here, as shown in Figure 44, is based on Bresnan (1996, 2001), which, in turn, follows Baker (1983).

The a-structure is a lexical syntactic representation with the minimally necessary information on the syntactic arguments of a predicator, such as the number of arguments, their thematic and syntactic types, and their hierarchical organi-

---

**Figure 42.**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chair BEI sit ASP</td>
<td>Floor BEI sleep ASP</td>
</tr>
<tr>
<td>‘The chair has been sat *(on).’</td>
<td>‘The floor has been slept *(on).’</td>
</tr>
</tbody>
</table>

**Figure 43.**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chair sit-threadbare ASP</td>
<td>Floor sleep-collapse ASP</td>
</tr>
<tr>
<td>‘The chair is threadbare from (over)sitting.’</td>
<td>‘The floor was slept on and collapsed.’</td>
</tr>
</tbody>
</table>
Optimality-Theoretic Lexical Mapping Theory

As shown in Figure 44, the a-structure “sink <ag pt>” states that the verb sink requires two arguments, one of the type agent and the other theme, and also that agent is thematically more prominent than theme. The a-structure thus contains information necessary for the final syntactic manifestation or more precisely, the mapping of agent and theme to SUBJ and OBJ, respectively.

Morpholexical operations interact specifically with lexical semantics and, as such, are outside the proper domain of the LMT, while morphosyntactic operations are part of the LMT, which constrain the syntactic assignment of a-structure roles. All OT-LMT constraints thus are conceived to be morphosyntactic and universal in nature, while morpholexical operations may be language-specific.

Potential Advantages of OT-LMT

The OT-LMT proposed here targets specifically at the universal constraints on argument-function linking. We leave the OT formulation of morpholexical operations to further research. In this section, we discuss some of the advantages that the OT-LMT may afford.

We start from the fact that all OT constraints are declarative. In the conventional LMT, the two subject conditions must apply sequentially, not simultaneously, to prevent $\theta[-r]$ from mapping to SUBJ, when $\theta[-o]$ is present. Likewise, only after the mapping of the subject role can other roles be mapped. In contrast, all constraints in OT-LMT apply declaratively and, thus, simultaneously. Furthermore, in the conventional LMT, all roles are uniformly assigned exactly one feature for func-
tion assignment, while the OT-LMT allows a more expressive system with only the patient/theme role pre-assigned to unrestricted functions. This OT-LMT thus allows the possibility of agent-OBJ as a marked morphosyntactic selection, which is ruled out in the conventional LMT.

Recall also that two disjunctions are observed in the conventional LMT: the disjunction between the two principles of subject role mapping and the disjunction between subject roles and non-subject roles. The OT-LMT, however, consistently favors the unmarked values for all roles. This characteristic ultimately may lead to the replacement of the stipulation in the Subject Condition while preserving its insight. Thus, in general, this revised LMT formulated in OT formalism offers a potentially more consistent and simpler computational system.

As noted earlier, the LinkLocInv constraint proposed in the OT-LMT account essentially reflects the insight of Bresnan and Kanerva’s (1989) locative default, which assigns \( \text{loc}_{-[r]} \) when \( \text{th} \) is focused. Thus, both accounts are descriptively equivalent in explicating locative inversion in the various languages observed, Chinese included. However, Bresnan and Kanerva’s (1989) account would need to state that languages like Japanese and Korean lack the mechanism of linking \( \text{loc} \) to \( [-r] \). The OT account, on the other hand, has the advantage of a more general solution in attributing the presence or absence of locative inversion in a language to the relative ranking of LinkLocInv, which, like all OT constraints, is universal.

Finally, we will indicate exactly how the OT-LMT better reflects the intuition that the locative inversion construction of \(<\text{th-OBJ loc-SUBJ}>\) is marked in comparison to the canonical locative construction of \(<\text{th-SUBJ loc-OBL}>\). In Figure 46, we examine the constraints that each of the two violates.

Notice that the canonical form constitutes only one violation of one of the two lowest-ranked constraints. The inverted form, on the other hand, violates two of the higher-ranked constraints, LinkRolRes and DescendFun, in addition to one of the two lowest-ranked constraints. Nonetheless, even with such violations, the inverted form still outranks all other candidates. It is, therefore, still the optimal choice, in spite of its markedness. The OT-LMT is, therefore, more expressive and flexible, accounts for a wider range of data, and reveals the (un)markedness of different linking relations. It is, in short, a simpler, more consistent, and more general theory.

**Directions of Further Research**

Considering its limited number of principles, LMT is a relatively small theory but with ambitious goals. In the previous sections, an OT version of the theory has been laid out and tested against cross-linguistic data of locative inversion as

---

**Figure 46.**

<table>
<thead>
<tr>
<th>Candidate</th>
<th>L.inkPrTh</th>
<th>LinkLocInv</th>
<th>L.inkRolRes</th>
<th>DescendFun</th>
<th>LinkUnobj</th>
<th>LinkUnres</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(&lt;\text{th-SUBJ loc-OBL}&gt;)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candidate</td>
<td>L.inkPrTh</td>
<td>LinkLocInv</td>
<td>L.inkRolRes</td>
<td>DescendFun</td>
<td>LinkUnobj</td>
<td>LinkUnres</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>------------</td>
<td>-------------</td>
<td>------------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>C4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(&lt;\text{th-OBJ loc-SUBJ}&gt;)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
generalized by Bresnan (1994). However, there are locative constructions closely related to locative inversion that have not been covered. For instance, locative inversion might bear some relation to sentences with an expletive subject and a locative argument (Bresnan, 1994). Figure 47 is an example from French.

The expletive subject is an athematic argument, and, as such, it must receive an intrinsic \([-r]\) classification by the very nature of thematic restrictedness \([r]\), (Bresnan, 2001). Given its initial position, it invariably links to \(\text{SUBJ}\)\(^7\). Similar to the locative inversion discussed earlier, this construction also indicates that the object function of the theme role renders it more focal than the oblique locative role. An athematic argument in the a-structure is indicated by an underscore outside of the angled brackets, while thematic arguments are within the angled brackets. Thus, the a-structure of Figure 47 is “\(\text{arrive} \_<\text{th}[\text{foc} \text{loc}]>\)” An expletive subject also may be associated with the linking of agentive objects. Figure 48 is an example from French.

Similar phenomena also are observed in Bantu languages (Demuth, 1990; Demuth & Mmusi, 1997; Harford, 1990; Machobane, 1995). All these issues are important and interesting but cannot be adequately addressed in the current article. Further research is needed on how the a-structure “\(\text{arrive} \_<\text{th}[\text{foc} \text{loc}]>\)” and “\(\text{work} \_<\text{ag}[\text{foc}]>\)” come about, whether they are morpholexically or morphosyntactically related to “\(\text{arrive} <\text{th loc}>\)” and “\(\text{work} _<\text{ag}[\text{foc}]>\),” respectively, and how best to incorporate such relations within the OT-LMT proposed here.

---

**Figure 47.**

Il est arrivé beaucoup de gens à la plage
it is arrived many of people at the beach
‘There were many people arriving at the beach.’

**Figure 48.**

Il travaille deux mille ouvriers dans cette usine
it works two thousand workers in this factory
‘There are two thousand workers working in the factory.’

**Figure 49. Asymmetrical object parameter (AOP)**

\[
\begin{array}{c|c|c|c|c}
* & \_\_ & \_\_ & \_\_ & \_\_ \\
\_r & \_r & \_r & \_r & \_r \\
\end{array}
\]
Further development of this OT-LMT also will need to address the issue of secondary patient-like roles as a parameter of variation in double object constructions, known as the Asymmetrical Object Parameter (AOP) (Alsina & Mchombo, 1993; Bresnan, 2001). In a non-AOP language, all patient-like roles are linked to an unrestricted function, while AOP languages must link the secondary patient/theme to an object function. An additional constraint may be necessary, and constraint ranking then may reflect this variation. This asymmetrical object parameter is stated in Figure 49.

Finally, the OT-LMT developed here needs to be applied to a much wider range of data cross-linguistically; for example, complex predicates in various languages (Abaitua, 1988; Ackerman, 1992; Alsina, Bresnan & Sells, 1997; Her, to appear; Ishikawa, 1985), the valence-changing morphemes and inversion constructions in Georgian (Blevins, 2005; Harris, 1981; Holisky, 1981), among others. A solid analysis of some of these facts would be a significant test of the linking theory proposed here.

CONCLUSION

In this article, we set out to accomplish two goals. The more ambitious one is to come up with a simpler and more general lexical mapping theory in OT terms, or OT-LMT. The second one is to test this theory and account for locative inversion in Chinese, English, and Chichewa on the one hand and Japanese and Korean on the other. Following the standard view in Optimality Theory, the mapping constraints we proposed are all universal, and language variation in locative inversion is accounted for by different constraint rankings. The OT-LMT we proposed is the UG component that constrains the argument-function linking, or morphosyntactic processes. It thus does not govern language-specific morpholexical processes, such as passivization, locative transitivity, and resultative compounding. Locative inversion, on the other hand, involves only morphosyntactic operations and, therefore, is accounted for within the OT-LMT.

In summary, the OT-LMT we proposed not only covers a wider range of empirical data, but it also affords a simpler, more consistent, and more general theory.

ACKNOWLEDGMENTS

I am genuinely grateful to the three anonymous reviewers for their insightful comments and exceedingly generous suggestions, and also to Adams Bodomo for his encouragement. I also thank Nissa Hui-shan Lin for discussions on OT. However, I am solely responsible for the content of the article. Research reported in this article has been funded partly by NSC grants 92-2411-H-004-024 and 93-2411-H-004-006. Part of the research for the article was done while I visited the School of Information Technology, Bond University in 2003 and 2004. I thank the dean, Professor Ron Davison, for his kind and continuous support.

REFERENCES


Maxwell, J. T., III, & Kaplan, R. M. (1996). An efficient parser for LFG. In M. Butt, & T. H. King (Eds.), *Proceedings of the LFG96 Conference* [online].


ENDNOTES

1 See Bresnan (2001, chapter 14) for a brief exposition of other formulations.

2 Falk (2001) also presents a concise introduction to LMT and a more precisely defined theory of argument roles. Dalrymple (2001) offers more examples in her introduction to the theory.

3 The concept of thematic hierarchy is well-established (Grimshaw, 1990; Li 1995). The hierarchy in “The Subject Condition” also might be derived from the proto-role properties proposed by Dowty (1991) (Ackerman & Moore, 2001b; Bresnan, 2001).

4 Note that following Zaenen and Engdahl (1994), the two propositional argument functions COMP and XCOMP are treated as instances of OBL.

5 Note that Θ refers to the most prominent role in the a-structure. The fact that it is also the left-most role within the angled brackets is inconsequential. Θ is usually also the initial role, unless there is an initial athematic argument.

6 Alsina (1996) also argues that the function-argument bi-uniqueness condition, which is fully integrated in the OT-LMT proposed later, is too strong.

7 A constraint for athematic roles must restrict such roles to [-r]. Alternatively, a constraint may be proposed to outrank LinkFun in Figure 22b and thus allow a GF in a-structure to be unmatched.

8 An additional constraint is needed for the secondary patient/theme, which restricts the secondary pt/th to [+o]. Again, we will ignore this issue in this article.

9 DescendFun and DescendRol effect a parallel alignment between the thematic hierarchy and the markedness hierarchy. This parallel alignment is similar to a harmonic alignment, but with an important difference. In a harmonic alignment, according to Prince and Smolensky (2004), the correspondence between a more prominent element on one scale and a less prominent element on the other is worse than the correspondence between two elements that are equal in prominence (Aissen, 1999; Asudeh, 2001; Lee, 2001; Sharma, 2001). Thus, it is better for agent, the most prominent role, to link to SUBJ, the most prominent function, and likewise for locative, the least prominent role, to link to OBJ. However, in our scheme here, due to LinkUnobj and LinkUnres, a more prominent GF is favored, regardless of the prominence of the role.

10 IC refers to the intrinsic classification of argument roles. See Figure 7.
Therefore, as I have proposed elsewhere (Her, 2003), morpholexical operations likewise can add features and, thus, alter syntactic assignments of argument roles, besides changing the lexical semantics of a predicator.

See Her (1989) and Ting (1998), among others, for compelling arguments against viewing the *bei*-NP phrase as a PP *by*-phrase. In fact, the same concept of suppression in passives is used here, as well. See Her (2004) for details of mapping the composite role, formed by two roles, to a single GF.

Note that I am only referring to computational efficiency in formulation and formalism, not in practical terms of an actual computational implementation. See Kuhn (2003) for extensive discussions on the computational aspects of OT. However, there is little practical evidence for the computational efficiency of a large-scale OT implementation of a grammar, as there seems to be no such practical systems yet. For the computational efficiency of LFG in general, see Maxwell and Kaplan (1996, 1993, 1991); for LFG in practice, refer to Kaplan, et al. (2004). I thank the anonymous reviewer who made this point and provided the references.

I thank another anonymous reviewer for pointing this out to me.

I thank the anonymous reviewer who suggested this direction of further research and provided this French example and its discourse analysis.

Refer to Bresnan (2001, section 14.1) for a more in-depth discussion on athematic arguments in raising constructions.
Section VIII
Emerging Trends

This section highlights research potential within the field of human-computer interaction while exploring uncharted areas of study for the advancement of the discipline. Chapters within this section highlight the socio-semantic web, podcasts, and the emergence of increasingly intelligent user interfaces. These contributions, which conclude this exhaustive, multi-volume set, provide emerging trends and suggestions for future research within this rapidly expanding discipline.
Chapter 8.1
Podcastia:
Imagining Communities of Pod-People

Jonathan Cohn
UCLA, USA

ABSTRACT

While podcasting has become a valuable advertising tool for many companies, it has also become a major way in which geographically spread out communities have been able to stay connected. Podcasts, like many other new Internet genres, are thought to be listened to mainly by an affluent audience who create podcast themselves. By looking at the various institutional and production issues and audiences of the podcast medium, this chapter will show how this genre works to create and sustain mass communities of “prosumers” and mobile audiences. Also, this chapter will historically contextualize the podcast by showing ways in which it is not simply a reiteration of earlier technologies, but also a distinct new medium with a unique, prosumer-friendly mode of transmission and reception.

INTRODUCTION

The sudden rise of podcasting on the “new media” scene has stunned, perplexed, and pleasantly surprised a vast majority of techno-savvy producers and consumers. According to the The New Oxford American Dictionary, which named “podcast” the word of the year in 2005 (“Wordsmiths hail,” 2005), this term refers to “a digital recording of a radio broadcast or similar program, made available on the Internet for downloading to a personal audio player” (“podcast n,” 2005). This is a simplified definition that does not take into account that almost any type of media can be podcasted, including audio, video, or even text. While “radio broadcasts” make up a large proportion of popular podcasts, they are by no means the only possible origins of a recording. While much of the reporting and advertising of podcasting has centered around commercial podcasts like recorded radio programs or newspaper articles, most are still made by amateurs who design them with the podcast as the intended platform. It has been most compared to amateur mediums like the communal ham radio, the subversive pirate radio, and the editorial and journalistic blog, which podcasts are usually formally attached to.

What separates a podcast from Internet radio, another genre of streaming media, which podcasts
Podcastia are often conflated with, is that a podcast is not live, and it is stored on the user’s computer or mp3 player. A podcaster records a show using any of a number of audio programs that can range from simple and free to professional and expensive. They then often use a sound-editing program, like the free Audacity or GarageBand, to edit and clean up their recording. These programs offer extensive Help Menus and excellent podcasting tutorials. The final steps involve getting your sound file online before they put it on their Web site attached to an RSS, or Really Simple Syndication feed. This is a file format that allows for media to be easily syndicated throughout the Web. There are many Web sites that will both turn your file into an RSS feed and upload the file to the Web for you. For example, professors are now beginning to record their lectures and put them on campus Web sites attached to an RSS feed. The RSS feed tells students when new recordings are available and allows them to download them automatically. For instance, Google search for “LectureCast” brings up many Web sites, like the popular www.podcastdirectory, which offer many free podcasts of UNLV lectures.

There are many different ways to create a podcast and finding the way that works best for you is largely based on your knowledge of sound editing and Internet technologies, as well as the type of computer and software you are using. Due to this large range of methods that can be used to achieve your goals, I will not be going into very much detail on the techniques of podcast creation, but luckily there are many great online tutorials that I advise everyone to explore. My favorite is www.how-to-podcast-tutorial.com. This site goes into a great deal of detail on both how to create and improve podcasts, as well as how to promote and think about your content. The Podcasting Tools Web site (http://www.podcasting-tools.com/how-to-podcast.htm) is another site that is particularly good at explaining how to record and edit audio. Most new audio recording software should also include a help section on podcasting as well. Also, a simple Internet search will bring up many more helpful sites and as this technology is still changing and growing, this is the best way to get up-to-date information on podcasting technology.

For podcast consumers, the process starts with them either scouring the Internet, or a podcast directory for LectureCast. There are many searchable directories online, but the most well known is the one attached to the iTunes Music Store (iTMS), which is part of Apple’s iTunes music software. After the professor’s podcast has been found, the user must subscribe, or download a link, to LectureCast into a “podcatching” program, like iPodder, or iTunes. This program will keep track of the LectureCast RSS link and download the individual lectures as soon as the professor attaches them to it. Consumers can then listen to them on their computers, or move them to an mp3 player so that they can listen to them on the go.

This mode of reception, which is not dependent on time or space, as most radio and television broadcasts are, has become particularly popular with those living in urban, subway-ridden environments, where radio transmissions are unreliable at best, and also those who have moved away from these cities, or have simply moved around often in general. They are also, like many other new Internet genres, thought to be listened to mainly by those who also produce them. By looking at the various institutional and production issues and audiences of the podcast medium, this chapter will show how this genre works to create and sustain mass communities of mobile audiences and “prosumers,” or consumers who also occasionally produce media. This term has long been used by electronics and camera equipment companies, who consider this market segment to be quite profitable.
BACKGROUND

The term, “podcasting,” was coined in The Guardian in February of 2004 (Hammersley, 2004), but it was not until an article in the March 2005 issue of Wired called, “Adam Curry Wants to Make You an iPod Rock Star,” that it moved from a very obscure and tech-centered genre, to a popular emerging genre. Curry, the former MTV VJ along with Dave Winer, the original author of RSS code, have both been major innovators and proponents of podcasting technologies since their inception. Curry is also credited with having produced one of the very first podcast, The Daily Source Code, which was a place “where developers and users party together” (2005). Even at its inception, podcasting was envisioned as a genre that could discard the distinctions between producers and consumers, as both groups were always part of the same creative community.

In the Wired article, Curry evangelizes this new genre, which he hopes will allow for amateurs to compete with commercial radio. He proceeds to make a podcast with his interviewer in the front seat of his car, while extolling the virtues of what he called “unregulated, low-cost, on-demand radio.” This forum has created, according to the interviewer, Annalee Newitz, products that are “obsessively internecine and gloriously, honestly unprofessional.” She also notes that everybody in the podosphere seems to know everybody else...On his show, Curry coos over a daily Bible podcast that features kids reading bits of the new Testament, while husband-and-wife podcast team Dawn and Drew speculate about whether Whole Wheat Radio is really recorded in a shack...When was the last time you heard your favorite morning show team crack an in-joke that referred to what a movie reviewer said two days ago on another station? (Newitz, 2005).

While in the television industry, until very recently, networks compete for viewers during timeslots, during this early period of podcasting, there was no competition, as the audience could listen to all of the podcasts that it wanted to in consecutive order. It also helped that there was no money involved, as this was a communal hobby for most people who “never would have had the time for Internet broadcasting if it hadn’t been as easy as “clicking a button and talking’” (Newitz, 2005). Newitz reports that many of the first podcasters took advantage of the easy, quick, and FCC-less production process by simply screaming expletives for long periods of time, playing indie songs they liked, lazily chatting on the couch, and giving sex advice. These interesting, if usually unsuccessful experimentations with the form only encouraged more listeners to become producers themselves as their seems to be no downside in expressing one’s amateurish self, or at least one’s guttural utterances of offensiveness, as loud as possible in this medium.

MAIN FOCUS

In 2005, podcasting became more popular and many of the products began to become more polished. One such podcast was Cinecast, a movie review show that featured two friends, Adam Kempenaar and Sam Hallgren, in Chicago. At the time of this book’s publication, this podcast is now called Filmspotting and can be subscribed to through the iTMS. Kempenaar saw the article in Wired and read the instructions in the article on how to cast. He saw how easy it was to do and called his friend Hallgren to ask, “Do you want to be a radio star?” (“Interview,” 2005). They heard all of the amateur-type shows and thought that although “that’s kind of the beauty of podcasting...We thought we could do something that seemed like a real radio show” (2005).
This has made them very popular with both the original podcasters community, and with those who were just discovering the technology. Curry even mentioned them in the first two episodes of his Podfinder podcast, “an audio directory” in which Curry points out “the latest and greatest in podcasting” (Curry, 2005). This is a very influential podcast among certain pod audiences and producers, and it is generally an honor to be included in it.

Cinecast was also featured on the front page by the iTMS when it started including podcasts on June 29, 2005. This resulted in a jump from hundreds to many thousands of listeners. This exposure, “along with a plug from podcast guru Adam Curry, sent their show skittering up to the iTunes list of the top 100 podcasts, peaking No. 13” (Green, 2005). Within a month, “Cinecast jumped from 250 to 50,000 subscribers” (Friess, 2005a). This large and sudden growth illustrates the importance of trusted gatekeepers, like Curry and iTunes, in emerging genres, as they tend to direct information traffic to what they consider worth at least checking out. By “gatekeeping,” I am referring to the way these entities process and filter information even as they relay it from the producer to the consumer. As gatekeepers, they decide which podcasts are heard and what information is important.

Curry fulfills his role as gatekeeper through both his Podfinder cast, and his larger network of podcasters, called Podshow network. Many networks, or communities of podcasters have formed in an effort to bolster their ranks, legitimate the format, and hopefully gain sponsorship (“PodShow,” 2006). This business model is based around the idea that by forming larger networks, advertisers will be able to feel more comfortable advertising through them. So far, placing commercials at the beginning or halfway through an episode has been one of the more popular podcasting business models for more established podcasters.

In Wired in September 2005, Tim Bourquin, the producer of Endurance Radio, a sports themed podcast, which was successful at gaining sponsorship from Gatorade and Fleet Feet Sports said that one of the biggest challenges was that “most potential sponsors don’t even know what a podcast is” (Friess, 2005a). By forming these larger communities, many podcasters have been able to gain a visibility and marketability that they probably could not have gotten on their own. This business model is similar to ad-based television, which is organized into networks, which sell a commodity, or the audience, to advertisers. However, it should be pointed out that the money podcasters are earning for their work is miniscule and virtually all independent podcasters have day jobs.

While Cinecast has Peerflix (a Netflix-like company) as their title sponsor, and other secondary sponsors (like HBO, USA Network and Honda) through Podtrac, their “network,” they aren’t actually making very much money: “we’ve been making enough off sponsorship and listener donations to cover our costs and pay a little bit out to ourselves…We’ve saved enough to start thinking about advertising, upgrading equipment, etc…But full-time job? No, we’re not really close to that at all” (A. Kempenaar, personal communication, June 6, 2006). It is certainly possible that as the podcast audience grows, the revenue stream will as well, but for now, most podcasters do it simply because they enjoy it and hope, or “expect profits to come later” (Friess, 2005b).

The addition of podcasts on iTMS, the most pervasive gatekeeper of all digital media, has also had a profound effect on the podcasting community. Most importantly, by legitimizing the technology and making it readily available to anyone who listens to, or watches media on their computer, iTunes made it extremely appealing to many television networks and film conglomerates. Now, only a handful of the iTunes top 100 podcasts are independently made, and those that are usually have an “explicit,” or “X,” rating
next to them. Many of the corporate podcasts are repurposed material from Comedy Central, Cartoon Network, ESPN, NPR, and various news channels and programs. Fox has podcasts for many of their shows, including Family Guy and The Simpsons, which are literally just the reading of a plot summary of the last week’s episode. Film studios also regularly now make podcasts to advertise their movies.

There is also now a large number of fan podcasts, devoted to discussing the worlds of Harry Potter, Lost and other popular programs. The quality of these fan podcasts ranges and many of these appear to be paid for by media conglomerates, but it is not always clear, as they don’t have commercials and they do not usually advertise their affiliations. Instead, the show itself acts as a commercial even as it acts as a site of a fan-based community in the same way sci-fi conventions do.

By looking at the iTMS top100 list, one might come to the conclusion that independent podcasts have been decimated by the onslaught of corporate media, but this is highly misleading and the situation is not so dire. There is a fear that “once a podcast drops off the top 100 list, it’s almost impossible for a casual visitor to find it...it will be harder for a new indie podcaster to get an audience” (Chin, 2005). It is true that when iTMS podcasting started, “47 of its most popular podcasts were produced by what some call “independents,” and by July 18, 2005, that number had dropped to 31 and continued to fall (Barnako, 2005), the actual audiences of most independent podcasts has risen considerably.

In one blog entry on the subject of whether iTunes was hurting indie podcasts, many people posted replies saying that even though their rankings might have gone down, their subscribers have greatly increased” (“Comments,” 2005-2006). Others also wrote in saying that their “subscriber…base has gone up 12x since our listing on iTunes. So being an independent “niche” podcast with no hope of breaking the top 100, we are thankful for iTunes” (2005-2006). While the very top podcasts are controlled by major companies, it is clear that there is a rather long, and large tail behind it filled with the independents, which can skew the total podcast listenership to the indies (Honan, 2005). iTMS also donates commercial banner space to some of the independent podcasters that they think should be noticed, like Cinecast, in the hopes of leveling the playing field a bit (A. Kempenaar, personal communication, May 12, 2006).

Users can also go to alternative, independent podcaster network and directory sites instead of iTunes to find more obscure titles that iTMS might not be carrying or promoting. There are many sites that are easily searchable by looking on Google, or by typing in addresses like www.podcast.com, which is one of many alternative sources for podcast feeds. Original, independent and amateur (if sometimes absolutely bizarre) podcasts still exist but they are just a bit more difficult to find than those made by Viacom and News Corporation.

The biggest dilemmas for some independent podcasters on iTMS come not when they fall off of the top 100 list, but when they suddenly find themselves on it. The day after podcasting went live on the iTMS, Cinecast, and many other casters suddenly discovered that the number of people accessing their podcast RSS feeds had exponentially increased, which meant they had to upgrade their server plan to one that allowed for more bandwidth. This can be expensive, as while a small, 100 Gigabyte server package can cost $25 a month, a 1,200 Gigabyte package, which was what was necessary for some to handle the increased traffic, is $300 a month, a much more expensive proposition for a group that mostly thinks of podcasting as a hobby that they do not expect to be compensated for (Friess, 2005a). Luckily, those that do make it onto this list usually pick up sponsors pretty quickly, which usually pay for at least the amount that it costs to get the podcast on the air.
However, this process of getting a sponsor can be extremely difficult for those who have only recently found their audience. Companies like to know the size and demographics of audiences before they commit to advertising and this information can be extremely difficult to figure out in this medium. iTMS doesn’t share these kinds of statistics with anyone and they also do not share how they rank podcasts on their top 100 and top 25 lists. This has irked many, as at times these two lists, which should be identical, have had unexplained discrepancies (Friess, 2005b). Theoretically, these issues should not be a problem when figuring out exactly how many subscribers a podcast has, as iTMS really only provides links to other Web sites and RSS feeds, and those other Web sites can normally keep a tally of how many people are using them.

This system breaks down, unfortunately, right as most independent podcasters need it to work best. If iTMS features a small podcast on one of their pages in the hopes of helping it find an audience, as they did with Cinecast, Apple will sometimes cache the podcast on their own server in order to make sure the added traffic doesn’t overwhelm, or crash the originating server. This is necessary, as iTMS “counted more than 1 million podcast subscriptions through iTunes in the first two days alone,” an impressive number considering that before this launch very few people even knew what a podcast was (Friess, 2005a). Unfortunately for the podcasters, this also means they have no way of knowing how many people are actually subscribing to their podcast through this program and cannot give accurate information to prospective advertisers.

Luckily, many businesses have sprung up in order to help podcasters with gathering information for advertisers. One such company is Podtrac, “The podcaster-advertiser connection” (www.podtrac.com, 2006). This entity provides surveys to podcasters, who put them on their Web sites, along with Third Party Measurements and Media Kits. For a cut, Podtrac will also help podcasters get advertising from companies like HBO, USA Network and Honda. Podtrac paid for a survey study in December 2005, that showed that only 32% of their correspondents were familiar with the term “podcasting,” 32% of those familiar (10-12% of the whole) had ever listened to one, and 41% of that group (4-5% of the whole) had listened to one in the last week (“Men are listening,” 2005). The study also found that 78% of those who had ever listened to podcasts were male, but women were just as likely to have listened in the past week. Podtrac expects this gender difference to disappear after a while and compares it to the early days of the Internet: In 1994, 95% of Internet users were male, but now that number is closer to 48%. They also point out that the PodTrac surveys for individual podcasts has shown that the demographics vary greatly from show to show, and therefore podcasts can act as a good outlet for advertisers to target specific audiences (2005).

While these numbers are small compared to most other mass medias, this ability to target specific audiences has convinced many that this could be a lucrative investment. Other surveys, from companies like Jupiter Research have come out with statistics saying that the average consumer of podcasts is “male, earns big bucks and...is a youngster” (Vaas, 2005). The Cinecast audience, according to their PodTrac survey of almost 700 people, is very similar, made up mainly of 18-34 year old white, college educated, single men (A. Kempenaar, personal communication, June 6, 2006). This has historically been a highly sought-after demographic and explains why many companies, like Chrysler, were “excited about supporting this new technology...in general, we’re finding that advertisers are looking for new and different ways to reach people” (Friess, 2005b).

Specifically, the group that can be targeted easily by podcasts in a way that is much more difficult for other types of media is the mobile audience. The mobile audience has been linked to all electronic modes of communication, from the radio to the television. In his essay, “The Technol-
ology and the Society,” Raymond Williams explores the “operative relationship between a new kind of expanded, mobile, and complex society and the development of a modern communications technology” (Williams, 2000). He explains that the television and radio complex “is characterized by two apparently paradoxical yet deeply connected tendencies of modern urban industrial living: on the one hand mobility, on the other hand the more apparently self-sufficient family home … a form of mobile privatization” (Williams, 2000). Williams uses the mobile audience as a useful way of thinking about how, in the 1950s, instantaneous media forms could unite and inform large populations. As it became easier to move around the country via freeways and live on the outskirts of cities in suburbs, television became a way in which these secluded and private homes could still be part of the imagined national community.

This same principle underlies the ways podcasting works to create and maintain spread out communities that are both national and global in size. Just as people turn on television sets to reestablish communication, or at least gather information about the national community, which they have separated themselves from through their mobile, privatized existence, iPod users listen to podcasts to reconnect to the outside world. The term “iPod” implies that by putting on your microphones, you are enclosing yourself inside of it, and removing yourself from whatever dirty bums have chosen to sit next to you on the subway. The ads for the iPod and iTunes both support this reading, as some of them show a woman walking down the street with earphones on, as her reflection dances to the music. Others show outlines of people with cool hair dancing in place as they listen to their iPods, as if the body’s movements were removed from the body itself. As a mobile medium, the iPod, and podcasts allow for you to virtually move wherever you might want to go, while your body stays in one place, “on the subway or waiting in line at the post office,” two of the more widely prescribed ways of listening.

Oddly enough, the place many mobile audiences seem to want to virtually go is home. WBEZ, the NPR affiliate in Chicago released many of their radio programs as podcasts in the hopes of exposing “listeners potentially around the world to the variety of the city.” They did get an international audience and e-mails from places like Australia and Iraq, but “a lot of them are displaced Chicagoans” (Coomey, 2006). This points to the popular use of podcasts to keep track of what is happening in their hometowns. There are many podcasts devoted to local news and many others that highlight various sports teams and regions. If one was a big West Virginia Mountaineers fan but had left the area, they could easily go on iTunes and find a podcast of local sports news. One can even listen to a podcast that will tell you the weather in Charleston, West Virginia and many other cities throughout America.

The ease of use and inexpensive nature of podcasting has also made it an easy way for families to stay up to date with each other over large distances. When Apple announced that they were going to make it easier to photocast (i.e., send collections of photos like one would send audio or video tracks using RSS feeds, to view on iPods or on your computer), they framed it as being an easy way for grandparents to send people in their family pictures they were taking. While it might seem just as straightforward to read an e-mail or look at a blog about your hometown or family, one does not always have the time to sit down at a computer and read through an entire blog or long e-mail. If one does not have the time for reading an e-mail, one also usually does not have the time to watch television and the podcast is an easy solution.

Even if one is traveling long distance in a car, by subway to work, or is running on a treadmill, the podcast allows one to stay firmly connected to local, national, and global news. Radio and other live broadcast technologies are not feasible in these scenarios as you either can’t get a clear signal or are moving so much that the signal can’t
stay strong. These are the kinds of activities that one is supposed to partake in while listening to podcasts, as emphasized by the glut of iPod to car adapters, clothing designed to hide iPods on “dangerous” subways, and arm bands to keep iPods firmly in place as you jostle about in time and space.

Conversely, the price of this ability to stay in place as you move around is that the technology could be said to dislodge you from your particular location in time and space and put you in the realm of the virtual. By necessity, podcasts are always taped in advance and it usually takes iTMS and other sites at least a day to make an episode available. Among other things, this means that if you are listening to an ESPN podcast, you are most likely hearing the announcers speculate about a game that has already taken place. Spatially, even if you are on a bus in Los Angeles, listening to a podcast about people starving in Sudan, you might no longer be aware of your surroundings and miss your stop.

In “The Poetics of Augmented Space,” Lev Manovich questions this image that people plugged into earphones, on the Internet, or running around in a virtual reality simulation are actually divorced from their physical bodies and positions in space. He points to the fact that the image of “a person checking her e-mail or making phone calls using her personal digital assistant/ cell phone combo while at the airport, on the street, in a car, or in any other actually existing space” has superseded the earlier image of a person with a plug in their head, wandering around cyberspace as their body lies limp and cold (2003).

Instead of removing you from space, augmented Reality “helps the user to do the work in a physical space by augmenting this space with additional information” (Manovich, 2003). If one thinks about the iPod as a way of augmenting, instead of obscuring space, the event of riding a bus while listening to a podcast about the poor in Sudan may become conjoined and the audio may affect what and how you see the people around you. It is also possible that by listening to your grandmother tell you in a podcast how her day was, the subway ride might become a little less “dangerous,” and a bit more like home.

These spatial and temporal connections and dislocations that can be made by listening to a podcast in one place that was made in another, are also created and sustained within the shows themselves. For example, while Cinecast started out as made up of a few sections, including recent film and DVD reviews, along with a weekly top 5 list (i.e., top 5 horror films, westerns, break up movies, etc.) quickly their episodes got longer and longer as they started to read listener’s e-mails and play their voicemail. While they started out ranging from 20-30 minutes long, they usually span 50-60 minutes and usually almost half of the podcast is dedicated to replying to e-mails. These e-mails range from asking how the host could have left one of their favorite movies off of their top 5 list, to correcting them for saying someone won an Oscar when they most certainly did not. All of the e-mails are treated as if they come from places of knowledge and usually the answer to why they left a movie out is because they haven’t actually seen it.

Every few weeks they also do short movie marathons, where they will watch a few films from a genre or director that they do not know enough about, and then talk about their reactions during the podcasts. The films they watch are always decided by listener’s e-mails and online responses, as are the choices of recent films to go see. Thus, while you listen to this podcast, you are constantly reminded that you are not the only one and that many others across the planet share your same interests and are just as knowledgeable as you, or the hosts of the podcast itself.

When referencing the e-mails, the hosts always say both the first name of the person and their physical location, which only makes this communal activity of bashing cinema and nostalgizing old films more explicit. Kempenaar said that while reading e-mails has not directly helped get them
sponsors, it has encouraged more people to start and continue to listen to the program. This is an old radio and television trick that helps to both create a larger, more active audience and also the image of said audience.

Anecdotally, Kempenaar said he was talking to someone at a company that puts content on phones and had obviously seen a lot of podcasts. He “provided 4 e-mails from a cross-section of listeners and they were so persuasive that he started listening to the show and is now a regular listener” (A. Kempenaar, personal communication, May 12, 2006). Cinecast, and many other podcasts, also have frappr maps. These are maps from http://www.frappr.com, that a podcaster, or anyone, can make a link to on their page. Then, anyone who wants to can go to that link, click on where on the map they are living, and see where all of the other listeners are from.

It is impossible to tell how many people are actually interested in this feature, but there are many “push-pins” on the page for Cinecast, spread out across the globe, though most of the people are located in urban areas along major bodies of water (as are, in general, most cineastes...and people). Kempenaar wrote that their acts of “promoting feedback is more to engage [their] audience and make them part of the discussion...the key is interesting feedback, not necessarily where it’s coming from. But it has been great to get e-mails from Germany, Sweden, France, Japan and so on...” (A. Kempenaar, personal communication, June 4, 2006). These techniques have helped make podcasts like Cinecast a large community of equals that stretches globally, in which everyone can learn from each other.

Unfortunately, this spread out audience also makes it more difficult to find and retain sponsors for these relatively small podcasts. Small and local businesses are usually the ones most likely to advertise on community-based programming, but since podcasting communities are usually spread so thin, there is no way for these types of establish-

ments to prosper from working in this medium. Cinecast’s first sponsor was ChicagoMixer.com, a dating service for the Chicago area. After seeing “very little increased traffic for its money,” the sponsor left and the show was later picked up by Peerflix, a service that is not very location specific, though they probably won’t make any money from the 20% of their listeners that are from “59 other countries” (Friess, 2005b).

**FUTURE TRENDS**

The virtues of podcasting and the communities that they can create are not simply good for amateurs, but also for radio, television, and film. While these other forms of media slowly appear to be infiltrating podcasts, some radio stations are beginning to incorporate podcasts into their programming. CurrentTV regularly features material that used to be or could become a podcast. The format of that channel is basically the same as a podcast network, but on a much larger scale.

In radio, the CBC recently said that the shows they have made into podcasts have helped reach to a new demographic and that they hope “podcasting will become a two-way street: ‘We hope to get content from average people who will effectively become citizen producers’” (Newitz, 2005). KYOU radio in San Francisco is now programmed almost completely out of podcasting programs. The hosts of Cinecast now have a monthly radio show on WBEZ in Chicago, and Adam Curry, the guru of podcastia, can be heard regularly on Sirius Satellite Radio. As amateur podcasts migrate into these more mature, or at least older, forms of media, podcasting will become a more prevalent way of experiencing world news, arts and culture. In the process, TV and radio are even now beginning to take a few lessons and are trying to go more global in their scope, and personal in their reach.
CONCLUSION

Podcasting has become one of the most popular ways for both amateur and commercial broadcasters to transmit their own programs. While the process of putting a podcast online is simple compared to getting a program on television or radio, it is notoriously difficult to sustain an audience for more than a short period of time. It is also difficult to make any money through this medium, as few will pay to listen to podcasts and advertisers still see the form as an uncertain platform with a small audience.

However, it has tremendous potential for use in classrooms and other educational settings. The largest dilemma for podcasters is the difficulty of making the public aware that their programs exist. This is not a problem for professors, who can simply announce to their classes how to find their programs. Professors can record their classes for students to listen to at home and augment their studying, conferences can record their panels for those who can’t attend, and libraries and other facilities can use them to announce new programs, archives, and upcoming events. This is a great way to advertise and increase the audience for local events as well as make these events known to a larger group of international listeners.

REFERENCES


Men are listening to podcasts but women are getting hooked, too. (2005, December 27). PodTrac Press Center. Retrieved February 8, 2008, from


**Key Terms**

**Autocasting:** The use of software to automatically create audio podcasts out of text on a blog or other Web site.

**Blog:** A shortening of Web Log. It is a diary, or log-like Web site. New entries are displayed at the top of the Web site and older ones get pushed down. Podcasts are often formally attached to blogs.

**Digital Audio Player:** Also commonly known as an MP3 player. These are often small handheld devices, which minimally contain memory storage, a battery, and audio output. Most, if not all, are able to play compressed MP3 music files, though MP3 is just one of many types of digitally encoded audio files. The most popular of these are iPods. While iPods have lent their name to “podcasts,” any type of digital audio player can play podcasts. Many cars and home entertainment systems now also play digital audio files.

**Gatekeeping:** The process of filtering and editing information for a consumer. This is done through the editing of information within a piece of media, or through the decision of which media will be advertised and made more easily available to the public.

**Internet Radio:** Audio that is streamed and listened to via the Internet. Unlike Podcasting, Internet radio is not saved to a user’s computer and can therefore not be listened to on an mp3 player or away from the computer itself.

**MP3:** Also known as MPEG-1 Audio Layer 3. This is a popular compressed digital audio format. Most podcasts are compressed as MP3s. While the sound quality of MP3s is not as good as an uncompressed audio file, most people find it to be adequate for their needs.

**Podcast:** A digital recording of audio, video, text or other media that is distributed by subscription over the Internet.

**Podcatcher:** Software on a user’s computer that manages podcasts. One can search for and download podcasts using a podcatcher. This program will then continually check online to see if any new files have been added to a subscribed podcast. If they have, then the Podcatcher will download them. The most common Podcatcher software is Apple’s iTunes Music Store.

**Prosumer:** Usually thought of as an amateur producer, or someone who normally consumes media, but occasionally also produces it. This
Podcastia

segment of the population has paradoxically been thought of as a lucrative market segment that in many ways has a great amount of independence from the mainstream economy.

RSS: Stands for Really Simple Syndication. This is a simple piece of XML code that allows online media to be easily and frequently updated and added to. A Podcast is a type of RSS feed.

This work was previously published in the Handbook of Research on Computer Mediated Communication, edited by S. Kelsey and K. St.Amant, pp. 679-689, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 8.2
Maximizing Web Accessibility Through User–Centered Interface Design

Soonhwa Seok
The University of Kansas, USA

ABSTRACT

Digital inclusion and Web accessibility are integral parts of modern culture and, as such, have implications for social accountability. The World Wide Web Consortium (W3C) has suggested standards and guidelines regarding the inclusion of people with special needs, with an emphasis on higher accessibility and adaptability as the main goal of Web design. The user interface is the place where users can interact with the information by using their minds. Users with special needs can acquire information by using a human centered user interface. This article highlights the need to investigate the relationship between cognition and user interface.

INTRODUCTION AND BACKGROUND

A highly developed digital society needs to be integrated in order to function to meet each individual’s differentiated needs. Inclusion is reflected in the modern special education classroom. This practice of teaching and learning mirrors the changes in society because the classroom is the place where a collaborative, collective and real-time community is built consisting of teachers, educational executives, administrators, researchers, and parents. The terms inclusion, zero reject, and universal design, characterize a fully inclusive classroom. Zero reject means “students cannot be excluded from educational services based on their disability, severity of disability, contagious conditions, or costs of services” (Wood, 2006, p.9). The term inclusion is used to designate educational services and placement of students
with disabilities in the general education classroom, facilitated by classroom-provided supports, adaptation, modification, and accommodation for each individual student (Hunt & Marshall, 2006; Wood, 2006). Finally, universal design for learning refers to a curriculum or educational approach using assistive or educational technology to individualize teaching and learning (Turnbull, Turnbull, & Wehmeyer, 2005). In the inclusive classroom, the learning characteristics of the student with the disability are analyzed, and all characteristics are integrated into a teaching unit that best meets the student’s needs.

DIGITAL INCLUSION AND WEB ACCESSIBILITY

Digital inclusion and Web accessibility are leading the change of the digital Web culture. This movement brings about zero reject and differentiated inclusion on the Web by using universally designed content with built-in flexibility. Digital inclusion involves “social inclusion, the ever-developing information and communication technologies,” and Web design for “equal accessibility and opportunity” for all the individuals, especially for the individuals with disabilities; Web accessibility is therefore, “social inclusion” (Bradbrook & Fisher, 2004, p. 2; World Wide Web Consortium [W3C], 2005, ¶ 5).

The ultimate goal of Web access is to improve productivity and the quality of communication. Productivity is improved as individual’s social function improves as a result of his or her efforts in a social unit. The term Web accessibility means that individuals with disabilities are able to function on the Web, including engaging in cognitive and physical activities, such as navigating, comprehending, analyzing, synthesizing, manipulating, producing and evaluating information. They can also communicate with others on the Web for their own purposes, and contribute to the Web culture. In short, Web accessibility results in productive and universal access to information and other individuals (Bowie, Adler, Civille, & Gill, 1996; W3C, 2005).

Web accessibility brings about social accountability for individuals who would otherwise be placed outside the Internet culture because of their physical, sensory, or cognitive limitations, as well as different cultural and linguistic backgrounds. The term disability is often not defined in a Web context, although it is broadly used in Section 508 of the American law, W3C, and other legal documents. This is because some individuals with special needs regard themselves as not having a disability and being very independent, although they may have functional limitations or barriers imposed by society and the environment (Hunt & Marshall, 2006; Thatcher, 2006; W3C, 2005). Multiple social attempts have been made to facilitate and build Web accessibility, including guidelines, standards, and Website evaluation efforts established by the Americans with Disabilities Act, Section 508, the W3C, the Center for Applied Special Technology (CAST), the British Broadcasting Corporation (BBC), IBM, and Microsoft (Thatcher, 2006).

GUIDELINES FOR WEB ACCESSIBILITY

The ADA is a piece of civil rights legislation that protects people with disabilities from discrimination in four major areas: (a) private-sector employment, (b) public services, (c) public accommodations, and (d) telecommunications. When passed in 1990, the ADA sought to reduce obstacles to “equal access to employment, state and local government programs, transportation, public buildings, and communication technology” (Wisdom, White, Goldsmith, Bielavitz, Davis, & Drum, 2006, p. 20). While the ADA attempts to include the Internet and other services, the law itself is not explicitly clear on the Internet access.
Although the accessibility of a Web site is ultimately the responsibility of the designer of the site, federal legislation sets the standard for accessibility. Section 508 was implemented to ensure that government Web sites are accessible; however, the law now extends to include all schools that receive federal and state funds (Carter, 2004). Meant to create high-quality information that is accessible to all viewers, key components of Section 508 state that there should be “a text equivalent for every nontext element; all information conveyed through color must also be conveyed without color; and frames should be titled with text that identifies the frame and facilitates navigation” (Jaeger, 2006, p. 170). By passing laws such as Section 508, the federal government has made it clear that Web accessibility is an important issue and must be provided in the field of education. Table 1 lists the URLs of organizations that have established Web accessibility guidelines.

W3C guidelines developed by Chisholm, Jacobs, and Vanderheiden (2000) identify the following ways to ensure Web content accessibility:

1. Provide equivalent alternatives to auditory and visual content.
2. Don’t rely on color alone.
3. Use markup and style sheets and do so properly.
5. Create tables that transform gracefully.
7. Ensure user control of time-sensitive content changes.
8. Ensure direct accessibility of embedded user interfaces.
10. Use interim solutions.
11. Use W3C technologies and guidelines.
12. Provide context and orientation information.
13. Provide clear navigation mechanisms.
14. Ensure that documents are clear and simple. (Chisholm, Jacobs, & Vanderheiden, 2000, ¶ 6).

Furthermore, Section 508-1194.22 provides Web accessibility guidelines related to 16 usability items (Section 508, 2006, § 1194.22):

1. A text equivalent for every nontext element shall be provided (e.g., via “alt,” “longdesc,” or in element content).
2. Equivalent alternatives for any multimedia presentation shall be synchronized with the presentation.
3. Web pages shall be designed so that all information conveyed with color is also

Table 1. Organizations with Web accessibility guidelines and their URLs

<table>
<thead>
<tr>
<th>Institution</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>British Broadcasting Corporation (BBC)</td>
<td><a href="http://www.bbc.co.uk/accessibility/">http://www.bbc.co.uk/accessibility/</a></td>
</tr>
<tr>
<td>U.S. Department of Justice (ADA)</td>
<td><a href="http://www.usdoj.gov/crt/ada/adahom1.htm">http://www.usdoj.gov/crt/ada/adahom1.htm</a></td>
</tr>
<tr>
<td>Center for Applied Special Technology (CAST)</td>
<td><a href="http://www.cast.org/">http://www.cast.org/</a></td>
</tr>
<tr>
<td>Microsoft Corporation</td>
<td><a href="http://www.microsoft.com/enable/">http://www.microsoft.com/enable/</a></td>
</tr>
<tr>
<td>U.S. Office of Governmentwide Policy (Section 508)</td>
<td><a href="http://www.section508.gov/">http://www.section508.gov/</a></td>
</tr>
</tbody>
</table>
available without color, for example from context or markup.

4. Documents shall be organized so they are readable without requiring an associated style sheet.

5. Redundant text links shall be provided for each active region of a server-side image map.

6. Client-side image maps shall be provided instead of server-side image maps except where the regions cannot be defined with an available geometric shape.

7. Row and column headers shall be identified for data tables.

8. Markup shall be used to associate data cells and header cells for data tables that have two or more logical levels of row or column headers.

9. Frames shall be titled with text that facilitates frame identification and navigation.

10. Pages shall be designed to avoid causing the screen to flicker with a frequency greater than 2 Hz and lower than 55 Hz.

11. A text-only page, with equivalent information or functionality, shall be provided to make a Web site comply with the provisions of this part, when compliance cannot be accomplished in any other way. The content of the text-only page shall be updated whenever the primary page changes.

12. When pages utilize scripting languages to display content, or to create interface elements, the information provided by the script shall be identified with functional text that can be read by assistive technology.

13. When a Web page requires that an applet, plug-in or other application be present on the client system to interpret page content, the page must provide a link to a plug-in or applet that complies with §1194.21(a) through (l).

14. When electronic forms are designed to be completed on-line, the form shall allow people using assistive technology to access the information, field elements, and functionality required for completion and submission of the form, including all directions and cues.

15. A method shall be provided that permits users to skip repetitive navigation links.

16. When a timed response is required, the user shall be alerted and given sufficient time to indicate more time is required (¶ 7).

Section 508 (2006, § 1194.24) also mandates specific modes of video and multimedia presentation. Three items were developed related to multimedia accessibility:

1. All training and informational video and multimedia productions which support the agency’s mission, regardless of format, that contain speech or other audio information necessary for the comprehension of the content, shall be open or closed captioned.

2. All training and informational video and multimedia productions which support the agency’s mission, regardless of format, that contain visual information necessary for the comprehension of the content, shall be audio described.

3. Display or presentation of alternate text presentation or audio descriptions shall be user-selectable unless permanent.

The Web-accessibility standards in W3C and Section 508 are designed to increase the usability, accessibility, and adaptability of the Web information system for people with sensory, cognitive, cultural, linguistic, and any other physical impairments, by means of the user-computer interaction, or user interface. The standards address the built-in, flexible technology or its application on the Web. Section 508 and W3C cover the following areas:

- Usability of software applications and operating systems
Maximizing Web Accessibility Through User-Centered Interface Design

- Accessibility of Web-based Internet information and application
- Application of built-in telecommunication product
- Flexibility of built-in video or multimedia product
- Assistive technology of self contained, closed products, desktop and portable computers.

The universal goal of a democratic society is to improve the quality of all citizens’ lives. Thus, Web accessibility should benefit all individuals, with or without disabilities. The Internet, especially for individuals with a disability, can provide information to improve the quality of all citizens’ lives.

DOMAINS OF INSTRUCTIONAL OBJECTIVES

Bloom, Engelhart, Frust, Hill, & Krathwohl (1956) defined the three domains of instructional objectives as cognitive, affective, and psychomotor domains. The three domains have been employed in instructional design, Web design, software design, and program design (Belanger, & Jordan, 2000; Bloom et al., 1956). Hodell (2000) added a fourth domain, the interpersonal domain. These domains respectively deal with (a) the process of the mind, (b) motor skills or physical movement, (c) individuals’ attitudes and feelings, and (d) interactivity between individuals (Belanger & Jordan, 2000; Bloom et al. 1956; Hodell, 2000).

All four domains should be considered when a Web environment is built, including three layers of the Web environment: (a) content, (b) user interface, and (c) infrastructure. These mirror the construct of Web design, the Website’s purpose, its usability, and the designer’s philosophy and plan.

Harris (1999) argued that each layer creates synergy between the layers, as well as between users and the layers. The content layer provides the user with knowledge, information, simulation, and other resources. The user interface layer called human computer interface or human-computer interaction allows the users to interact with the content (information), the computer itself, and other users. Finally, the infrastructure layer, which includes the inner, hardware, mechanical system and technology of the machine, supports the interface, content, the user, as well as interactions between all of them.

COGNITION AND THE USER INTERFACE

Web accessibility is a construct that reflects digital inclusion, focusing on individuals with special needs and facilitating human-centered interaction and functionality. The user interface is a computer, Web or technology system that enables users to see, hear, and interact with the content, the computer, or other users. Thus, the user interface is a collective representation of cognitive, psychomotor, affective, and interpersonal processes that can be used to enhance the usability of the system. All users, including those with disabilities, need effective ways to access and acquire knowledge. An effective interface allows individuals to use their minds without heavy physical or mental efforts, which especially benefits individuals with various special needs. This highlights the need to investigate the relationship between cognition and user interface. The next section will discuss cognitive processing models.

COGNITIVE PROCESSING MODELS

The cognitive processing models developed by cognitive constructivists emphasize the most effective strategies for building knowledge (Jonassen, Davidson, Collins, Campbell, & Hagg, 1995; Miller & Miller, 1999; Wilson, Teslow, &
Osman-Jouchoux, 1995). Thus, the cognitive processing model considers knowledge as reality and an obtainable objective outside of the individual. Therefore, the individual and knowledge are independent of each other, at a distance, in their own structure. Further, within this model, knowledge is objectively measurable so that the individual can “acquire, replicate, and retain an accurate representation” (Jonassen et al., 1995; Miller & Miller, 1999, ¶ 18; Wilson et al., 1995). The model provides directions for acquiring information.

Cognitive constructivists employ three types of memory to acquire correct information: sensory, short-term, and long-term. Sensory memory is used to emphasize the concentration of the mind for a short time, while information is processed. Strategies in this area include psychophysical variation (e.g., change in color, size, pitch) and discrepant activities (e.g., choice of words, images)” (Miller & Miller, 1999, ¶ 21-23). When the strategies of iconic short term memory are implemented in designing Website for individuals with disabilities, items can be divided into (a) four to five chunks (Cattell, 1886), and (b) visual or audible partial content or cue representations, instead of whole reports (Averbach & Sperling, 1961). These strategies are generally employed in teaching practices for all students. Finally, strategies employing models of long-term memory include (a) activation of prior knowledge, (b) presenting information within a meaningful context, and (c) hierarchical sequencing and organization that elaborates and connects new with existing knowledge (Miller & Miller, 1999, ¶ 21-23).

Appropriate strategies for acquiring information should be applied to authentic, functional experiences that have a purpose in the context of the individual’s life. Strategies should also encourage the individual to pay attention to the information, so the experience can be transferred to sensory store, short-term memory, and long-term memory with little cognitive load. Figure 1 illustrates the information processing system and types of memory (Ally, 2004, p. 9).

The construct of Web accessibility can more precisely and diversely form this structure (Figure 1) than any other application because the Web construct is adaptable and flexible. The needs of the individual user and the function of the individualized Website determine the structure of the display, degree of user control of navigation, and level of interactivity. The essential goal of this cognitive constructivist paradigm for Web accessibility is easy and accurate transmission and reception of functional information. Digital inclusion generates synergy by applying the cognitive processing models to the individual’s needs, and builds an authentic digital functional community, which is the nature of the Web environment.

---

**Figure 1. Types of memory**

![Diagram of information processing system and types of memory](image-url)
THE HUMAN-COMPUTER INTERFACE

The human-computer interface (HCI) is a crucial element in producing meaningful and determining factors of usability (Kim & Moon, 1998; Nielsen & Norman, 2000; Roy, Dewit, & Aubert, 2001). Specifically, the user interface is a computer system (Ciavarelli, 2003) that connects the contents or infrastructure and Web users. Thus, users can operate, control, and manage the contents and infrastructure through the medium and user interface (Ciavarelli, 2003; Harris, 1999; Laurillard, 1993). Further, the user interface “provides feedback to the user, such as acknowledging inputs, advising the user of processing wait periods, and providing navigational assistance, error alerts, and corresponding recovery methods” (Ciavarelli, 2003, p. 14).

The goal of the interface is to produce higher usability that will increase users’ effectiveness and efficiency in using the computer and navigating its contents or texts. It also enhances users’ satisfaction and perception of the interaction with the computer (Ciavarelli, 2003; Harris, 1999; Nielsen, 1993; Roy et al.). Higher usability is reflected by the following characteristics: (a) requiring minimum efforts to operate, (b) visually agreeable and enjoyable to use, and (c) pursuing few errors and quick recovery from them (Howe, 2005). Therefore, usability is a “cognitive, social,” communicational process with an integrated computer technology system (Adler & Winograd, 1992; Berg, 2000, p. 353; Maddix, 1990).

HCI refers to “the design, evaluation, and implementation of interactive computing systems for human use” (Berg, 2000, p. 350; Card, Moran, & Newell, 1983; Ciavarelli, 2003; Head, 1999). Therefore, HCI is the set of knowledge about interaction between the computer systems, users, and productive working environment (Berg, 2000; Carey, 1991; Ciavarelli; Reisner, 1987). Tasks related to theories support interface design, including (a) the perceptual approach (Gibson, 1986); (b) constructivism (Ciavarelli; Duffy & Jonassen, 1992); and (c) the activity theory or information processing approach.

THE GOMS MODEL

Based on the cognitive information processing model, Card, Moran, and Newell (1983) developed a cognitive model of human computer interaction, GOMS. As discussed earlier, the cognitive information processing model has three different memory stages; (a) sensory store, (b) short-term memory, and (c) long-term memory. The sensory store and working memory stages are related to perceptual and motor processing, whereas short-term memory and long-term memory are related to cognitive processing (Card et al., 1983). According to Card et al. (1983), the GOMS model consists of four components that demonstrate how to process tasks between users and a computer system: (a) goals to be achieved in the process, (b) operators performing the sequencing methods and interacting with the system, (c) methods representing the sequences of the tasks performed by the operators, and (d) selection rules choosing the best strategies and solutions for the best methods to obtain the goals. The GOMS model seeks to “predict the time required to complete the task. In addition, the model can be used to identify and predict the effects of errors on task performance. Error recovery is assumed to involve the same four components as correct actions” (Card, Moran & Newell, 2005. ¶ 2).

ERROR AND COGNITION

Error types mirror the working load of short-term memory. Error occurs when the short-term memory has a heavy workload. There seem to be millions of errors in everyday life. Some research
shows that errors have several visible patterns, which can be prevented or reduced by the human centered interface design. James Reason and Donald Norman are two of the most notable psychologists with expertise in error models as illustrated in their books, Human Error and The Design of Everyday Things, respectively. In addition, they deal with the effective use of the human mind and the design of objects that we encounter everyday.

Reason (1990) used Rasmussen’s (1986) classification of human performance to generate his generic error modeling system (GEMS), which is the classic error model associating human errors with the actions or action planning of the working memory. The GEMS model involves three types of action control: skill, rule, and knowledge-based levels. Reason argued that slips and lapse are errors that result from skills, whereas mistakes result from knowledge.

Reason (1990) linked error types to the cognitive steps, planning, storage, and execution, those for which short-term memory is mostly in charge. According to Reason (1990) human error can be categorized into three observable, measurable behaviors: skill, rule, and knowledge:

At the skill-based level, human performance is governed by stored patterns of preprogrammed instructions represented as analogue structures in a time-space domain. Errors at this level are related to the intrinsic variability of force, space or time coordination. (p.43)

Reason asserted that at the rule-based level, “errors are typically associated with the misclassification of situations leading to the application of the wrong rule or with the incorrect recall of procedures” (p. 43). At the knowledge-based level, errors “arise from resource limitations (‘bounded rationality’) and incomplete or incorrect knowledge” (p. 43). As expertise increases, “the primary focus of control moves from the knowledge-based towards the skill-based levels; but all three levels can co-exist at any one time” (p. 43).

Reason (1990) claimed that the knowledge that is meaningful in a specific context or is not correct, is the main cause of errors. Thus, he argued that most errors are related to the knowledge level, whereas the two skill-level errors are slips and lapses. Based on Reason’s error frame, slips are errors caused by over-attention or inattention after an action is completed while intention is forming and the execution or storage of action sequence is failed.

According to Norman, slips are “the performance of an action that was not what was intended” (2002, p. 1). Thus, Norman argued that slips appear as results of the similar, triggering actions or thoughts with inattention. Norman’s natural observation resulted in six categories of slips: “capture errors, description errors, date-driven errors, associative activation errors, loss-of-activation errors, and mode errors” (p. 107). He employed an activation-trigger schema system to explain schemata (2002). Rumelhart defined schemata as “data structures for representing generic concepts stored in memory” (Rumelhart & Ortony, 1977, p. 101); “schemata have variables with constraints and that these variables (or slots) have a limited distribution of possible default values” (Reason, 1990, p. 35; Rumelhart & Ortony, 1977). In Norman’s an activation-trigger schema system, errors mostly appear when schema is appropriately chosen and activated but triggered and operated over the longer time frame as a response to activators. This activation-trigger schema system is schematic control model using the working short-term memory; it handles familiar knowledge with speed without heavy workloads (Reason, 1996).

The activation framework has both specific and general activators. The specific activator uses a mental model which is composed of explanations of intended actions and plans (Norman, 2002; Reason, 1996). A less detailed explanation for the planned action is necessary if a collection of intended actions is to be effectively carried
Maximizing Web Accessibility Through User-Centered Interface Design

at short intervals. General activators support the work of the mind by providing contextual and affective background information, regardless of intentions.

Human-centered interface design analyzes and synthesizes an individual’s needs, cultural practice, cognitive ability, and preferences against the measurable goal of functioning in a certain context, using cognitive, psychomotor, and affective constructs (Badre, 2002). Applied to human center design, the systems increase usability, visibility, and functionality: “They save lives” (Norman, 2002, p. iv) by reducing errors of every day life. The outcome of the human-centered design is “ease of use, ease of learning, memorability, lack of errors, and satisfaction” (Barde, 2002, p.5; Gould & Lewis, 1985; Nielsen, 1993).

When the design that enhances human-computer interaction is considered in the context of Web accessibility, analysis and synthesis of the individual’s special needs and the work of the mind, “human perception, memory, and attention,” is important so as to ensure the minimal use of their efforts and interaction with technology (Norman, 2002, p. iv). The interdependence and interaction between people, information, and technology exists with a special function and purpose (where there is communication). Thus, analyzing and defining the user’s characteristics is the first step to building a Website for individuals with special needs. Their interactions with technology, content, information and other individuals are to be synthesized in the user interface.

Understanding users, their needs, mastery level of technology and their physical and cognitive ability is a prerequisite for designing the human computer interface and evaluating Web accessibility. The goal is to determine the most appropriate interaction modes and reduce the rate of error by providing physical, semantic, cultural and logical constraints (Badre, 2002; Norman, 2002).

Ecology (Web accessibility) should be adapted to meet the needs of individuals in the different categories of disability and the different degrees within a category (severe to mild). Barde (2002) described individual factors that influence Web design. Individual differences may be “grouped into four categories: knowledge, experience, and skill; personality factors; physical and demographic attributes; and user levels” (p. 70). Individuals may differ in “level and type of education as well as in their knowledge, experience, and skill levels” (p. 70). Barde elaborated:

These factors describe the cognitive abilities and styles of projected users as well as their knowledge and experience of the projected Website’s domain. Not all factors are necessarily applicable in every Web design situation, but there will be situations in which they all have direct implications for designing the constituents of the interactive environment (p. 70).

HUMAN-CENTERED DESIGN

Design is communication. Human-centered design yields synergy between the content information, the user, the designer, and infrastructure, and enhances communication effectiveness. Norman (2002) suggested that human-centered design consists of three elements:

1. Conceptual models which make invisible functions visible by using feedback, as the effect of an interaction, and explanations of its use.
2. Constraints which are a proactive measure to limit the choices of interaction and reduce human errors. Physical, semantic, and cultural constraints are three types of constraints.
3. Affordance is the “perception and actual properties of the thing. It suggests how the device can be possibly operated (p. 9).
The three elements make up a psychology of communicative design. Communication enhances visibility and usability by using the communication medium and human-computer interaction. Usability is defined as “the degree to which users can perform a set of required tasks” (Brinck, Gergle, & Wood, 2002, p. 2). The usability of Web accessibility is measured by the degree of efficiency of the individualized Web design and communication with the technology: functionality given the individual’s needs; efficiency of operating time, ease in remembering and learning operations as determined by the individual’s physical, cognitive ability and the degree of disability; proactive error-tolerant system; and built-in flexibility of the infrastructure design for the digital inclusion (Brinck, Gergle, & Wood, 2002; Norman, 2002).

**DISABILITY CATEGORIES**

Most special education textbooks divide disabilities into 12 categories, in accordance with the Individuals with Disabilities Education Act (see Table 2). The potential population of the Web accessibility users reaches over six million, age 6 to 21. Table 2 also illustrates the number of potential Web users in each disability category (Turnbull, 2004).

Section 508 and the W3C do not clearly define Web accessibility users in terms of the individual’s age or the degree of disability, only recommending that the Web be made “accessible to people with disabilities” (W3C, ¶ 1). The W3C indicates how the Web may be employed by users with different types of disabilities and “how people with disabilities use the Web.” The W3C

---

**Table 2. Categories of disabilities and corresponding numbers and percentages of students ages 6 to 21 served in the 2003-2004 school year**

<table>
<thead>
<tr>
<th>Disability</th>
<th>Number</th>
<th>Percentage of total disability population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special learning disabilities</td>
<td>2,858,260</td>
<td>47.4</td>
</tr>
<tr>
<td>Speech or language impairments</td>
<td>1,127,551</td>
<td>18.7</td>
</tr>
<tr>
<td>Mental retardation</td>
<td>581,706</td>
<td>9.6</td>
</tr>
<tr>
<td>Emotional disturbance</td>
<td>483,850</td>
<td>8.0</td>
</tr>
<tr>
<td>Multiple disabilities</td>
<td>132,333</td>
<td>2.2</td>
</tr>
<tr>
<td>Hearing impairments</td>
<td>71,903</td>
<td>1.2</td>
</tr>
<tr>
<td>Orthopedic impairments</td>
<td>68,188</td>
<td>1.1</td>
</tr>
<tr>
<td>Other health impairments</td>
<td>452,045</td>
<td>7.5</td>
</tr>
<tr>
<td>Autism</td>
<td>140,920</td>
<td>2.3</td>
</tr>
<tr>
<td>Visual impairments</td>
<td>25,814</td>
<td>0.43</td>
</tr>
<tr>
<td>Traumatic brain injury</td>
<td>25,534</td>
<td>0.37</td>
</tr>
<tr>
<td>Developmental delay</td>
<td>65,921</td>
<td>1.09</td>
</tr>
<tr>
<td>Deaf-blindness</td>
<td>1,667</td>
<td>0.03</td>
</tr>
</tbody>
</table>
developed series of scenarios of Web usability by individuals with disabilities: (a) an individual taking online courses with hearing impairment, (b) an accountant with visual impairment, (c) a student with a specific learning disability, and (d) a student with a cognitive disability transitioning from high school. In each case, the assistive technology and the human-computer interface reduced the work-load of short-term memory to enhance the usability of the given Website. The technologies and human-computer interactions are as follows:

1. Keyboard equivalents for mouse-driven commands;
2. Captioned audio portions of multimedia files
3. Appropriate markup of tables, alternative text, abbreviations, and acronyms; synchronization of visual, speech, and Braille display
4. Use of supplemental graphics; freezing animated graphics; multiple search options
5. Magnification; stopping scrolling text; avoiding pop-up windows
6. Clear and simple language; consistent design
7. Consistent navigation options
8. Multiple search options (W3C, ¶ 2)

CONCLUSION AND RECOMMENDATIONS

Web accessibility, especially human-centered, is an ideology because it involves an effort to change the traditional Web culture into digital inclusion, whereby the Web design meets an individual’s special needs considering the user’s social, cultural, physical, and cognitive characteristics. Access to the Web translates into access to knowledge and information that enable people to work in a digital society, and therefore, be digitally included. Access to information in turn, translates into access to society and the world. The Web must be accessible to all individuals, regardless of any physical or intellectual disabilities one may have. In order to ensure Web accessibility, guidelines such as the W3C standards and federal legislation such as Section 508 of the Rehabilitation Act have been developed.

It is important to research and design a human-computer interface that will enhance the usability and effectiveness of communication between the users, and between users and technology. Based on the error and cognitive processing models presented here, cognitive, psychomotor, and affective domains should be considered when the user interface is developed. As illustrated, the activities of action planning and storage take place at the knowledge and skill level in the cognitive domain. Execution occurs based on judgment, experience or preference. Execution is related to the affective domain. One of the most effective ways to reduce error is to present conceptual models/frameworks (Nielsen, 1993; Jonassen, Strobel, & Gottdenker, 2005), which enable Web users to construct knowledge of how to navigate, determine what can be done, and the sequence of work, with little cognitive demand. Thus, conceptual models help Web users explain and interpret (a) how to access information and technology, (b) how to adapt the context of the interface, and (c) how to communicate with the computer, itself. Jonassen et al. (1995) argued that users can build cognitive strategies to reason, engage, support, and assess information using the qualitative and quantitative model.

Web accessibility must be differentiated and individualized to meet the individual’s special needs. Developing effective interfaces for individuals with disabilities involves (a) identifying the user’s purpose of digital activity, b) identifying the information and placing it sequentially and, (c) if possible, applying Bloom’s five learning objective levels: knowledge, comprehension, analysis, synthesis and evaluation. This classification scheme will improve communication
between the information and technology. The relationship between processing and acquiring information is better-linked, and the information is more efficiently retrieved. The levels will provide insight into information processing and individuals with disabilities may interact with the technology simply and easily. In addition, strategies using short-term memory have been adapted from memory strategies for exceptional children and youth, such as (a) chunking information by dividing it into multiple pages, (b) repeating patterned rules to navigate a site, (c) representing meaningful, functional information, and (d) categorizing of information (Hunt & Marshall, p.128). Sound or visual effects, such as feedback or cues at certain intervals, will enhance attention by making hard-to-view items visible.

In a digital age, technology is always present—any where, any time. Winner (1980) argued that technology structures the mode of communication, work and consumption in a society. Technology is a pattern of modern culture; it is not just technological device (Borgmann, 1984). Intertwined with cultural, societal and economic factors, the Web is a subculture of modern technology. It is a structure by which we communicate and connect with the external world. In today’s digital information society, information is knowledge and can be used as a tool with which people can work (Bowen, Bereiter, & Scardamalia, 1992). The Web is a major source for acquiring information and connecting with the world, especially for the people with physical and intellectual limitations; all without the limitations of time and space. Therefore, all individuals need access to the Web, thus allowing them more opportunities to access and process information independently (Gilbert & Driscoll, 2002).

To meet the needs of our modern technological culture, the number of Web sites continues to increase. With this increase, concerns regarding the quality of Web sites and Web accessibility arise. A major concern is the need for evaluation strategies of Web site effectiveness, and usability, and a system of moral principles or ethics. Other concerns involve (a) the importance of knowing what factors of Web designs maximize Web accessibility, (b) how best to structure content and information for Web delivery and accessibility, and (c) how to evaluate accessibility in a manner that is valid and reliable, while also providing evidence on how to improve the quality of Web information. Given the growth in Web accessibility and the varied circumstances under which it is offered, evaluation instruments and processes are needed to ensure societal, cultural, educational, technical, and economic equity of high-quality Web accessibility.

REFERENCES


Maximizing Web Accessibility Through User-Centered Interface Design


KEY TERMS

Americans with Disabilities Act (ADA): A civil rights legislation that protects people with disabilities from discrimination against them in four major areas: 1) private-sector employment, 2) public services, 3) public accommodations, and 4) telecommunications.

Cognitive Processing Models: Models developed by cognitive constructivists emphasizing effective strategies for building knowledge.

Domains of Instructional Objectives: Cognitive, psychomotor, affective and interpersonal domains addressing (a) the process of the mind, (b) motor skills or physical movement, (c) individuals’ attitudes and feelings, and (d) interactivity between individuals.

GOMS Model: A cognitive model of human computer interaction consisting of four components: (a) Goals to be achieved in the process, (b) Operators performing the sequencing methods and interacting with the system, (c) Methods representing the sequences of the tasks performed by the operators, and (d) Selection rules choosing the best strategies and solutions for the best methods to obtain the goals.

Human-Centered Interface Design: A system used to analyze and synthesize an individual’s needs, cultural practice, cognitive ability, and preference with the measurable goal and function in a certain context, using cognitive, psychomotor, and affective constructs.

Section 508: A law implemented to ensure that government Web sites are accessible. The law now extends to include all schools that are provided with federal and state funds.

Web Accessibility: Equal accessibility and opportunity to the Web for all the individuals, especially for individuals with disabilities.

This work was previously published in the Handbook of Research on Web Information Systems Quality, edited by C. Calero, M. Moraga, and M. Piattini, pp. 206-219, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 8.3
Voice-Enabled User Interfaces for Mobile Devices

Louise E. Moser
University of California, Santa Barbara, USA

P. M. Melliar-Smith
University of California, Santa Barbara, USA

ABSTRACT

The use of a voice interface, along with textual, graphical, video, tactile, and audio interfaces, can improve the experience of the user of a mobile device. Many applications can benefit from voice input and output on a mobile device, including applications that provide travel directions, weather information, restaurant and hotel reservations, appointments and reminders, voice mail, and e-mail. We have developed a prototype system for a mobile device that supports client-side, voice-enabled applications. In fact, the prototype supports multimodal interactions but, here, we focus on voice interaction. The prototype includes six voice-enabled applications and a program manager that manages the applications. In this chapter we describe the prototype, including design issues that we faced, and evaluation methods that we employed in developing a voice-enabled user interface for a mobile device.

INTRODUCTION

Mobile devices, such as cell phones and personal digital assistants (PDAs), are inherently small, and lack an intuitive and natural user interface. The small keyboards and displays of mobile devices make it difficult for the user to use even the simplest of applications. Pen input is available on PDAs, but is difficult to use on handheld devices. Voice input and output for mobile devices with small screens and keyboards, and for hands- and eyes-free operation, can make the user’s interaction with a mobile device more user friendly. Voice input and output can also facilitate the use of Web Services (Booth, Hass, McCabe, Newcomer, Champion, Ferris, & Orchard, 2004) from a mobile device, making it possible to access the Web anytime and anywhere, whether at work, at home, or on the move. Global positioning system (GPS) technology (U.S. Census Bureau, 2006)
Voice-Enabled User Interfaces for Mobile Devices

can provide location information automatically for location-aware services.

Many everyday applications can benefit from voice-enabled user interfaces for a mobile device. Voice input and voice output for a mobile device are particularly useful for:

- Booking theater and sports tickets, making restaurant and hotel reservations, and carrying out banking and other financial transactions
- Accessing airline arrival and departure information, weather and traffic conditions, maps and directions for theaters, restaurants, gas stations, banks, and hotels, and the latest news and sports scores
- Maintaining personal calendars; contact lists with names, addresses, and telephone numbers; to-do lists; and shopping lists
- Communicating with other people via voice mail, e-mail, short message service (SMS), and multimedia message service (MMS).

It is important to provide several modes of interaction, so that the user can use the most appropriate mode, depending on the application and the situation. The prototype system that we have developed supports client-side, voice-enabled applications on a mobile device. Even though the applications support multimodal input, allowing keyboard and pen input, we focus, in this chapter, on voice input and on multimodal output in the form of voice, text, and graphics. The prototype includes a program manager that manages the application programs, and six voice-enabled applications, namely, contacts, location, weather, shopping, stocks, and appointments and reminders.

BACKGROUND

A multimodal interface for a mobile device integrates textual, graphical, video, tactile, speech, and/or other audio interfaces in the mobile device (Hjelm, 2000; Oviatt & Cohen, 2000). With multiple ways for a user to interact with the applications, interactions with the device become more natural and the user experience is improved. Voice is becoming an increasingly important mode of interaction, because it allows eyes- and hands-free operation. It is essential for simplifying and expanding the use of handheld mobile devices. Voice has the ability to enable mobile communication, mobile collaboration, and mobile commerce (Sarker & Wells, 2003), and is becoming an important means of managing mobile devices (Grasso, Ebert, & Finin, 1998; Kondratova, 2005).

The increasing popularity of, and technological advancements in, mobile phones and PDAs, primarily mobile phones, is leading to the development of applications to fulfill expanding user needs. The short message service (SMS) is available on most mobile phones today, and some mobile phones provide support for the multimedia messaging service (MMS) to exchange photos and videos (Le Bodic, 2002). The mobile phone manufacturers are no longer focused on making a mobile phone but, rather, on producing a mobile device that combines phone capabilities with the power of a handheld PC. They recognize that the numeric keypad and the small screen, common to mobile phones of the past, do not carry over well to handheld PCs (Holtzblatt, 2005).

With the emergence of Web Services technology (Booth et al., 2004), the Web now provides services, rather than only data as it did in the past. Of the various Web Services available to mobile users today, the map application seems to be the most popular, with online map services available from Google (2006) and Yahoo! (2006b). Much progress has been made in creating the multimodal Web, which allows not only keyboard and mouse navigation but also voice input and output (Frost, 2005).

GPS technology (U.S. Census Bureau, 2006) already exists on many mobile devices, and can be used to provide location-aware services (Rao
Voice-Enabled User Interfaces for Mobile Devices

& Minakakis, 2003), without requiring the user to input geographical coordinates, again contributing to user friendliness.

Speech recognition technology (Rabiner & Juang, 1993) has been developed over many years, and is now very good. Other researchers (Kondratova, 2004; Srinivasan & Brown, 2002) have discussed the usability and effectiveness of a combination of speech and mobility. Currently, handheld voice-enabled applications use short commands that are translated into functional or navigational operations. As observed in Deng and Huang (2004), speech recognition technology must be robust and accurate, and close to human ability, to make its widespread use a reality. Noisy environments present a particular challenge for the use of speech recognition technology on mobile devices and, therefore, multimodal interactions are essential. For example, the MiPad system (Deng, Wang, Acero, Hon, Droppo, Boulis, et al., 2002; Huang, Acero, Chelba, Deng, Droppo, Duchene, Goodman, et al., 2001) uses a strategy where the user first taps a “tap & talk” button on the device and then talks to the device.

Distributed speech recognition (Deng, et al., 2002), in which the speech recognition happens at a remote server exploits the power of the server to achieve fast and accurate speech recognition. However, studies (Zhang, He, Chow, Yang, & Su, 2000) have shown that low-bandwidth connections to the server result in significant degradation of speech recognition quality. In contrast, local speech recognition (Deligne, Dharianpragada, Gopinath, Maison, Olsen, & Printz, 2002; Varga, Aalburg, Andressy, Astrov, Bauer, Beaugeant, Geissler, & Hoge, 2002) utilizes speech recognition technology on the mobile device, and eliminates the need for high-speed communication. Local speech recognition limits the kinds of client handsets that are powerful enough to perform complicated speech processing and, thus, that can be used; however, the computing power of mobile handsets is increasing.

THE PROTOTYPE

The prototype that we have developed allows mobile applications to interact with the user without the need for manual interaction on the part of the human. Speech recognition and speech synthesis software are located on the mobile device, and make the interaction with the human more user friendly. The prototype that we have developed processes natural language sentences and provides useful services while interacting with the user in an intuitive and natural manner. A user need not form a request in a particular rigid format in order for the applications to understand what the user means.

For our prototype, we have developed six application programs and a Program Manager. These applications are Contacts, Location, Weather, Shopping, Stocks, and Appointments and Reminders applications. The Program Manager evaluates sentence fragments from the user's request, determines which application should process the request, and forwards the request to the appropriate application.

The prototype is designed to interact with a human, using voice as the primary means of input (keyboard, stylus, and mouse are also available but are less convenient to use) and with voice, text, and graphics as the means of output. The speech recognizer handles the user's voice input, and both the speech synthesizer and the display are used for output. Characteristics of certain applications render a pure voice solution infeasible. For example, it is impossible to convey the detailed contents of a map through voice output. However, voice output is ideal when it is inconvenient or impossible for the user to maintain visual contact with the display of the mobile device, and it is possible to convey information to the user in that mode. Voice output is also appropriate when the device requests confirmation from the user.

Thus, an appropriate choice of speech recognition and speech synthesis technology is vital to
Voice-Enabled User Interfaces for Mobile Devices

the success of our prototype. Our choices were constrained by:

- The processing and memory capabilities of typical mobile devices
- The need for adaptability to different users and to noisy environments

The use of speech recognition and speech synthesis technology on a mobile device is different from its use in call centers, because a mobile device is associated with a single user and can learn to understand that particular user.

The Underlying Speech Technology

The prototype uses SRI’s DynaSpeak speech recognition software (SRI, 2006) and AT&T’s Natural Voices speech synthesis software (AT&T, 2006). It currently runs on a handheld computer, the OQO device (OQO, 2006). We chose this device, rather than a cell phone, because it provides a better software development environment than a cell phone.

Speech Recognition

The DynaSpeak speech recognition engine (SRI, 2006) is a small-footprint, high-accuracy, speaker-independent speech recognition engine. It is based on a statistical language model that is suitable for natural language dialog applications. It includes speaker adaptation to increase recognition accuracy for individuals with different accents or tone pitches. It can be configured so that it performs speech recognition specific to a particular individual. DynaSpeak is ideal for handheld mobile devices, because of its small footprint (less than 2 MB of memory) and its low computing requirements (66 MHz Intel x86 or 200 MHz Strong Arm processor).

DynaSpeak supports multiple languages, adapts to different accents, and does not require training prior to use. It incorporates a Hidden Markov Model (HMM) (Rabiner & Juang, 1993). In an HMM, a spoken expression is detected as a sequence of phonemes with a probability associated with each phoneme. A probability is also associated with each pair of phonemes, that is, the probability that the first phoneme of the pair is followed by the second phoneme in natural speech. As a sequence of phonemes is processed, the probability of each successive phoneme is combined with the transition probabilities provided by the HMM. If the probability of a path through the HMM is substantially greater than that of any other path, the speech recognizer recognizes the spoken expression with a high level of confidence. When the response is below an acceptable confidence threshold, the software seeks confirmation from the user or asks the user questions.

The HMM is augmented with grammars for the particular applications that are required for understanding natural language sentences (Knight, Gorrell, Rayner, Milward, Koeling, & Lewin, 2001). When the user says a new word, the word can be added to the vocabulary dynamically. The HMM is also extended by adapting the vocabulary of the speech recognizer to the current and recent past context of interactions of the user with the applications.

Accuracy of the speech recognition system can be increased by training it for the voice of the particular user. There are two kinds of training, explicit and implicit. Explicit training requires the user to read a lengthy script to the device, a process that is likely to be unpopular with users. Implicit training allows the device to learn to understand better its particular user during normal use. Implicit training can be provided in two modes, confirmation mode and standard mode.

In confirmation mode, the system responds to a user’s sentence, and the user confirms or corrects the response. If the user corrects the sentence, the learning algorithm tries to match a rejected, lower probability, interpretation of the original sentence with the user’s corrected intent. If a match is found, the learning algorithm adjusts
the HMM transition probabilities to increase the probability of selecting the user’s intent. Initially, a new user of the system will probably prefer confirmation mode.

In standard mode, the system does not confirm sentences for which there is one interpretation that has a much higher probability than any other interpretation. If no interpretation has a high probability, or if several interpretations have similar probabilities, the speech recognition system responds as in confirmation mode. More experienced users of the system are likely to use standard mode.

The success of implicit training strategies depends quite heavily on starting with a speech recognizer that is well matched to the individual speaker. It is possible, from relatively few sentences, to classify a speaker and then to download, to the mobile device, an appropriate initial recognizer for subsequent implicit training.

DynaSpeak can be used with either a finite-state grammar or a free-form grammar. We used the finite-state grammar because it offers greater control over parsed sentences. The tendency for DynaSpeak to accept or reject spoken sentences is heavily influenced by the complexity of the grammar. The complexity of the grammar is quantified by the number of paths by which an accepting state can be reached. The greater the complexity of the grammar, the higher its tendency to accept an invalid spoken request. Conversely, the lower the complexity of the grammar, the higher its tendency to reject a valid spoken request. To minimize the complexity of the grammar and to improve speech recognition accuracy, each application has its own relatively simple grammar. The program manager determines which applications are involved in a sentence and then reparses the sentence using the appropriate grammars.

Speech Synthesis

Natural Voices (AT&T, 2006) is a speech synthesis engine that provides a simple and efficient way of producing natural (rather than electronic) sounding device-to-human voice interactions. It can accurately and naturally pronounce words and speak in sentences that are clear and easy to understand, without the feeling that it is a computer that is speaking.

Natural Voices supports many languages, male and female voices, and the VoiceXML, SAPI, and JSAPI interface standards. Using Natural Voices, we created text-to-speech software for our prototype that runs in the background and accepts messages in VoiceXML format. Each message contains the name of the voice engine (i.e., “Mike” for a male voice and “Crystal” for a female voice) and the corresponding text to speak.

Managed Applications

For the prototype we developed six multimodal applications (contacts, location, weather, shopping, stocks, appointments, and reminders) that use speech as the main form of input. The stocks, maps, and weather applications exploit existing Web Services on the Internet. Communication with those Web Services uses a local WiFi 802.11 wireless network. The program manager controls the operation of the applications. The graphical user interface for the program manager with the six applications is shown in Figure 1. We now present an explanation of the functionality of each application and its role in the overall system.

Contacts

The contacts application stores personal information regarding friends and acquaintances in a database, including their addresses and phone numbers. The contacts application is a mobile extension of a physical contact list or address book that is controlled by voice input. It retrieves data from Microsoft Office Outlook® to populate the database when in docking mode. After using the mobile device and possibly entering new contact information, the user can synchronize informa-
tion on the mobile device with that on a desktop or server computer. The contacts application is configured to interact with other applications that require information about names, addresses, phone numbers, and so forth. The contacts grammar is the least complex of the application grammars that we developed. The contacts vocabulary grows linearly as contacts are added to the user’s contact list.

Location

The Location application allows the user to search for restaurants, movie theaters, banks, and so forth, in a given area, using the Yahoo! LocalSearch Web Service (2006b). For example, if the user says to the mobile device “Search for a Mexican restaurant in 95131,” the location application on the mobile device sends a Web Service request to Yahoo! LocalSearch, gets back the results, and presents up to 10 results to the user in list form. The user can then view additional information about a single location by indicating the location’s number in the presented list. For example, the user can choose to view additional information about “Chacho’s Mexican Restaurant” by speaking, “Get more information about number one.” On processing this request, the location application presents the user with detailed information about the restaurant including its phone number, address, and a detailed street map showing its location. Figure 2 shows a screen shot of the graphical user interface for the location application.

The location application is loosely coupled with the contacts application to provide responses related to individuals listed in the user’s contact list. For example, the request, “Search for a movie theater around Susan’s house” uses the contacts grammar to determine the location of Susan’s house and replaces the phrase “Susan’s house” with the specific address so that the actual search request looks something like this: “Search for a movie theater around 232 Kings Way, Goleta, CA, 93117.” The location application then searches for a movie theater in the vicinity of that address.

The location application is also loosely coupled with a GPS module that is contacted when the user has a question related to the user’s current location. For example, if the user says “Look for a pizza place around here,” the word “here” is recognized by the application and replaced with the GPS coordinates of the user’s current location.
The location application then sends a Web Service request to Yahoo! LocalSearch, which returns a map of the user's current location, indicating where the user is, along with the 10 nearest pizza places. The Yahoo! LocalSearch Web Service is ideal to use with GPS because of its ability to locate positions on the map on the basis of longitude and latitude. With GPS, the user is no longer limited to requests involving a particular city or zip code. The user now has the ability to create requests that are truly location-aware.

Compared to the grammars of the other applications, the location grammar is one of the most complex. For information like maps and lists, it is desirable to use a graphical or textual display, as well as speech output, in a multimodal user interface. Thus, the most appropriate kind of output can be chosen, depending on the kind of information, the capabilities of the mobile device, and the context in which the user finds himself or herself.

Weather

The weather application supplies weather forecasts obtained from the Web Service provided by the National Weather Service (NOAA, 2006). It allows the user to query for weekly, daily, and 3-day weather information in major U.S. cities using voice input. It allows the user either to select a city or to use the user's current location, as the location for which the weather forecast is to be retrieved from the National Weather Service. The weather application knows the geographical coordinates of dozens of cities in the continental United States. It references those coordinates when the user requests a weather forecast from the National Weather Service for one of those cities.

A user can say “Tell me the weather forecast in San Jose,” which then uses “today” as the starting time of the forecast, and produces the graphical user interface for the weather application shown in Figure 3.

Because the weather application operates on a mobile device, it is necessary to be able to determine the user’s location dynamically. If the user asks “What’s the weather like here two days from now?” the weather application consults the GPS module to obtain the geographical coordinates of the user, contacts the Web Service, and responds with the high and low predicted temperatures and an indication that there is a change to cloudy in
Thus, the user does not need to provide his/her current location or to obtain the weather forecast for that location.

Our prototype takes into account the many ways in which a person can convey, semantically, equivalent requests in English. For example, a user can ask for the weather in many ways including “What is the weather in Boston like?” or “Tell me what the forecast is like in Boston.” These two requests are semantically equivalent because they both contain the same essential parameter, namely the Boston location.

Shopping

The shopping application provides the user with a service capable of reducing the time that the user spends on grocery shopping and the associated stress. The shopping application maintains shopping lists, recipes, and floor plans of supermarkets. The multimodal interface includes speech, text, and graphics, which makes the shopping application easy to use. Figure 4 shows a screen shot of the graphical user interface for the shopping application.

The shopping application allows a user to update his/her shopping list and to forward it to another user. When a user issues a command, like “Remind John to go grocery shopping,” the contacts application is used to find John’s phone number or e-mail address in the user’s contact list. A dialog box then appears asking the user if he/she wants to send, to John, not only a reminder to go shopping but also the shopping list. If so, the shopping list, consisting of the product ids and the quantities of the items needed, is formatted in XML, and appended to the message containing the reminder. The message is then sent to John’s shopping application.

The shopping application also displays graphically the floor plan of the supermarket and the location of items in the store, as shown in Figure 4. This feature provides assistance to the user without the need for the user to contact an employee of the supermarket. The shopping application also allows the user to retrieve recipes while shopping.
possibly on impulse, for an item that is on sale. A newly chosen recipe is cross-referenced with the current shopping list, so that needed items can be added automatically. The shopping application has the largest grammar of the applications that we developed, with a vocabulary that depends on the items that the user has purchased recently.

**Stocks**

The stocks application allows the user to manage his/her stock portfolio using voice input and output. The objective of the stocks application is to monitor stock fluctuations, rather than to trade stocks. The stocks application exploits the Yahoo! Finance Web service (2006a) to store and update stock information in a database. It stores the most recent stock information in the database so that it can reply to the user’s requests when connectivity to the Yahoo! Finance Web Service is limited. Although such stored data can be somewhat stale, it allows the user to obtain information whenever the user requests it. The vocabulary of the stocks application grows to match the user’s portfolio each time the user adds a new stock.

**Appointments and Reminders**

The appointments and reminders application manages the user’s calendar and allows the user to send reminders to other people. It supports time-based requests of various forms, for example, “Remind me to go to the dentist on Monday,” “Remind me to see the dentist on August 15th,” and “Remind me to see the dentist a week from today.” It displays an easily readable schedule, so that the user can recall what is planned for the day. The appointments and reminders application interacts with other applications, such as the shopping application. For example, the request “Remind John to go shopping on Monday” sends a reminder to John, along with the current shopping list, if the user wishes to forward that information. It also supports reminders to the user that are location-aware using GPS, for example, if the user is in the vicinity of a supermarket. The appointments and reminders...
application is an extension of a calendar service. It links to Microsoft Office Outlook®, and updates scheduled appointments and reminders when in the vicinity of the user’s desktop.

**Program Manager**

The program manager evaluates sentence fragments from a user’s request, identifies keywords that determine which application or applications should process the request, reparses the sentence using the grammars for those applications, and forwards the parsed request to the appropriate application. If more than one user is involved, the program manager on one user’s mobile device sends messages to the program manager on another user’s mobile device, which then handles the request.

The program manager leverages DynaSpeak and a weighted keyword recognition algorithm to break down recognized sentences into application-specific fragments. Those fragments are then processed by the appropriate applications, and are subsequently merged to form the final sentence meaning. This process allows the program manager to handle requests that involve more than one application, for example, “Search for a gas station around Paul Green’s house.” The parsing of this sentence, using the location grammar, requests a search centered on a location that the location grammar cannot itself provide. The program manager must recognize a keyword from the contacts grammar, parse the sentence using that grammar, and query the contacts application for the address of Paul Green’s house. The response to the query is then sent to the location application to obtain the location of the gas station nearest his house.

**Graphical User Interface**

The graphical user interface (GUI) of the program manager, shown in Figure 1, displays the current running application programs and allows the user to select an application by using voice or keyboard input. The GUI provides buttons that appear gray when an application has not been started and blue after startup. If the user makes a spoken request
that requires an application to display a result, the display for that application is topmost and remains topmost until the user issues another request or a timeout occurs. Whenever the GUI is displayed, the user must provide a keyword in a spoken request to wake up the program manager, or click on one of the application-specific buttons on the display.

**EVALUATION**

Several experiments were performed to collect qualitative and quantitative data to evaluate the prototype system. Although it is difficult to determine a clear boundary between the user interface and the speech recognizer, it is important to evaluate the user interface and the speech recognizer separately, so that the qualitative and quantitative data gathered from the experiments are not mixed, leading to inconclusive results.

Thus, the experiments were designed as a classical “Don't mind the man behind the curtain” study. In this type of study, the user interacts with a system that is identical to the actual system except that the experiment is being controlled by someone other than the user. The man behind the curtain controls what is spoken as responses to the user’s requests and changes the current screen to an appropriate graphical response. This method was used, so that the responses to the qualitative questions would not be biased by the accuracy of the speech recognizer.

To evaluate the system quantitatively, the program manager was instrumented with time segment metrics and data were collected for several performance metrics, including:

- Total time a participant took to complete all tasks
- Overhead of the DynaSpeak speech recognizer during live and batch recognition
- Runtime overhead of the program manager without DynaSpeak
- Spoken length of a request vs. processing time

The results are shown in Figure 6. The time segment metrics represent the runtime complexity of the code associated with the speech recognition.

*Figure 6. Processing overhead per task*
and processing. The amount of time taken by each segment adds to the delay associated with the user’s request. If any of the time segments has a large duration, the user might become irritated. By measuring each segment separately, the bottleneck in the system can be determined.

The speech processing time increases with the size of the grammar. However, by means of a multiphase procedure that uses keywords organized and weighted by application relevance, the grammar size and the speech processing time can be improved. After live recognition, the system provides a keyword-associated request, which it processes for application weights and then reprocesses using an application-specific grammar, possibly more than once with different grammars. This procedure increases both the speed and the accuracy of the speech recognition, by decreasing the size of the grammar size in the initial phase.

An alternative approach (Kondratova, 2004) is to force the user to make repeated requests, possibly from a menu, with responses by which the device asks for the next step or for more information, so that the device arrives at a better understanding of the user’s request. Such an approach introduces navigational complexity for the user. Reducing the speech processing time by creating a complex navigational structure is not the best way to improve usability of the system.

The speech recognizer works better for some speakers than for other speakers. The accuracy of the results can be improved by tuning the speech recognition parameters and enabling learning capabilities. However, the developers of DynaSpeak advise against modification of the speech recognition parameters and use of learning until a relatively high success rate is achieved. For appropriately selected users, quite good speech recognition and understanding can be achieved without using learning capabilities. However, speech recognition accuracy can only improve if voice profiling is combined with learning.

Ambient noise and microphone quality also affect speech recognition accuracy. The internal microphone in the OQO device is of rather poor quality. To ameliorate this problem, a Jabra© Bluetooth headset, was used to provide noise cancellation and reduce the distance between the microphone and the user’s mouth. In addition, when the confidence score from DynaSpeak falls below an acceptable threshold, the program manager seeks confirmation from the user or asks for clarification. These mechanisms greatly improve the accuracy of the speech recognizer.

The accuracy of speech recognition is degraded when the grammar contains words that are phonetically similar. During preliminary experiments for the shopping application, we had problems recognizing differences between similar sounding requests like “Add lamb to my shopping list” and “Add ham to my shopping list.” These problems arise particularly when users are non-native English speakers or when they have accents. Creating more specific requests can reduce the phonetic similarity, for example, by saying “Add a lamb shank to my shopping list” and “Add a ham hock to my shopping list.” However, modifying requests in such a way is undesirable because the requests are then less intuitive and natural.

The location, weather, and stocks applications all use Web Services and require communication over the Internet and, thus, have longer application runtimes than the other Web Services. The location application is written in Java, which runs more slowly than C#. Both the weather application and the stocks application cache data associated with previous requests to take advantage of timing locality. Location requests are different because the caching of maps can involve a large usage of the memory, and users are not inclined to perform the same search twice. Memory is a precious commodity on a handheld device and needs to be conserved; thus, the location application is coded so that it does not cache maps resulting from previous queries.
To evaluate the qualitative aspects of the system, we performed a user study with participants from diverse backgrounds of education, ethnicity, and sex. The user study was completed with 10 individuals performing 10 tasks resulting in 100 request results. The participants were given a questionnaire that assessed their general impressions about the prototype, with the results shown in Table 1.

After analyzing the averaged responses of the participants, we found several trends. The participants’ scores are not strongly correlated with speech recognition accuracy. Participant G gave the system a high score, but was one of the two participants who encountered the most speech recognition problems. Participant B gave the system a low score despite good speech recognition.

The participants agreed that speaking to a mobile handheld device as if it were a human is not comfortable. It is difficult to get used to interacting with a computer that can understand tasks that would be commonplace for humans. The participants were relatively pleased with the GUI interface design and felt the system is relatively easy to use. However, the ease-of-use metric needs to be taken lightly. Ease of use can be assessed more concretely by measuring the number of times a user must repeat a command.

The scores for response appropriateness and relevance are high, indicating that the spoken responses of the applications were well crafted. The scores related to recommending the service to friends and daily life helpfulness are relatively high, from which one might infer that the participants would purchase a device providing the speech-enabled applications. However, this conclusion is not necessarily justified. The participants were not enthusiastic about having to pay for such a device or for such services. However, most participants in the study were quite pleased with the prototype system and found the user interface helpful and easy to use.

**Table 1. Responses to the questionnaire**

<table>
<thead>
<tr>
<th>Questions</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was it comfortable talking to the device as if it were a human?</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>3.7</td>
</tr>
<tr>
<td>Was the GUI aesthetically pleasing?</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>Were the request responses appropriate and easy to understand?</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4.2</td>
</tr>
<tr>
<td>Were the spoken responses relevant to your requests?</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4.5</td>
</tr>
<tr>
<td>Was the system easy to use?</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4.3</td>
</tr>
<tr>
<td>Do you think the services would be helpful in your daily life?</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4.2</td>
</tr>
<tr>
<td>Would you recommend a system like this to your friends?</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4.2</td>
</tr>
<tr>
<td>Would you buy the software if it were available for your phone?</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>3.8</td>
</tr>
</tbody>
</table>
grammed to work together, as in our prototype. However, future systems will need to support tens or hundreds of applications, many of which will be designed and programmed independently. Integration of those applications and their grammars will be a challenge.

Currently, speech-enabled applications typically use short commands from the human that are translated into navigational or functional operations. More appropriate is speech recognition technology that supports a more natural, conversational style similar to what humans use to communicate with each other (McTear, 2002).

A mobile device that listens to its owner continuously can provide additional services, such as populating the user’s calendar. For example, when a user agrees to an appointment during a conversation with another person, the mobile device might recognize and automatically record the appointment, possibly confirming the appointment later with its user. Similarly, the mobile device might note that the user habitually goes to lunch with the gang at noon on Mondays, or that the user leaves work promptly at 5pm on Fridays. With existing calendar systems, the user often does not record appointments and other commitments, because it is too much bother using the human interfaces of those systems, greatly reducing the value of the calendar.

A useful capability of speech recognition systems for mobile devices is being able to recognize intonation and emotional overtones. “The bus leaves at 6” is, overtly, a simple declaration, but appropriate intonation might convert that declaration into a question or an expression of disapproval. Existing speech recognition systems do not yet recognize and exploit intonation. Similarly, the ability to recognize emotional overtones of impatience, uncertainty, surprise, pleasure, anger, and so forth, is a valuable capability that existing speech recognition systems do not yet provide.

Speech recognition requires a relatively powerful processor. Typical cell phones contain a powerful digital signal processor (DSP) chip and a much less powerful control processor. The control processor operates continuously to maintain communication with the cellular base stations. The DSP processor uses a lot of power and imposes a significant drain on the battery and, thus, analyzes and encodes speech only during calls. The DSP processor is capable of the processing required for speech recognition, although it might need more memory.

For mobile devices, battery life is a problem, particularly when speech recognition or application software requires a powerful processor. The limit of 2 hours of talk time for a cell phone is caused at least as much by the power drain of the DSP processor as by the power needed for wireless transmission. The DSP processor might be needed for speech processing for more than 2 hours per day. There are several possible solutions to this problem, namely, larger batteries, alcohol fuel cells, and DSP processors with higher speeds, reduced power consumption, and better power management.

Background noise remains a problem for speech recognition systems for mobile devices, particularly in noisy environments. The quality of the microphone, and the use of a headset to decrease the distance between the microphone and the speaker’s mouth, can improve speech recognition accuracy.

**CONCLUSION**

The use of voice input and output, in addition to text and graphics and other kinds of audio, video, and tactile interfaces, provides substantial benefits for the users of mobile devices. Such multimodal interfaces allow individuals to access information, applications, and services from their mobile devices more easily. A user no longer has to put up with the annoyances of a 3-inch keyboard, nested menus, or handwriting recognition, nor
does the user need to have a tethered desktop or server computer in order to access information, applications, and services. Providing multiple ways in which the users can interact with the applications on mobile devices brings a new level of convenience to the users of those devices.

REFERENCES


Voice-Enabled User Interfaces for Mobile Devices


**KEY TERMS**

**Global Positioning System (GPS):** A system that is used to obtain geographical coordinates, which includes a GPS satellite and a GPS receiver.

**Hidden Markov Model (HMM):** A technique, based on a finite state machine that associates probabilities with phonemes, and pairs of phonemes, that is used in speech recognition systems, to determine the likelihood of an expression spoken by a user of that system.

**Location Aware:** An application that is based on a particular physical location, as given by geographical coordinates, physical address, zip code, and so forth, that determines the output of the application.

**Mobile Device:** For the purposes of this chapter, a handheld device, such as a cell phone or personal digital assistant (PDA), that has an embedded computer and that the user can carry around.

**Multimodal Interface:** The integration of textual, graphical, video, tactile, speech, and other audio interfaces through the use of mouse, stylus, fingers, keyboard, display, camera, microphone, and/or GPS.

**Speech Recognition:** The process of interpreting human speech for transcription or as a method of interacting with a computer or a mobile device, using a source of speech input, such as a microphone.

**Speech Synthesis:** The artificial production of human speech. Speech synthesis technology is also called text-to-speech technology in reference to its ability to convert text into speech.
**Web Service:** A software application identified by a Uniform Resource Indicator (URI) that is defined, described, and discovered using the eXtensible Markup Language (XML) and that supports direct interactions with other software applications using XML-based messages via an Internet protocol.

*This work was previously published in the Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 446-460, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 8.4
Multilayered Approach to Evaluate Mobile User Interfaces

Maria de Fátima Queiroz Vieira Turnell
Universidade Federal de Campina Grande (UFCG), Brazil

José Eustáquio Rangel de Queiroz
Universidade Federal de Campina Grande (UFCG), Brazil

Danilo de Sousa Ferreira
Universidade Federal de Campina Grande (UFCG), Brazil

ABSTRACT

This chapter presents a method for the evaluation of user interfaces for mobile applications. The method is based upon an approach that combines user opinion, standard conformity assessment, and user performance measurement. It focuses on the evaluation settings and techniques employed in the evaluation process, while offering a comparison between the laboratory evaluation and field evaluation approaches. The method’s presentation and the evaluation comparison will be supported by a discussion of the results obtained from the method’s application to a case study involving a Personal Digital Assistant (PDA). This chapter argues that the experience gained from evaluating conventional user interfaces can be applied to the world of mobile technology.

INTRODUCTION

As proposals for new techniques and methods emerge for the evaluation of mobile device usability, it becomes more difficult for practitioners to choose among them. To be able to evaluate the efficacy of these techniques and methods, as well as to reproduce their steps, they have to be described in a level of detail not often found in the literature. Claims are often made without solid statistical results and are usually based on superficial descriptions. This makes it difficult, if not impossible, to compare alternative choices. Given the features of these new devices (such as mobility, restrictive resources for information input and output, and dynamic contexts of use), HCI specialists may question the efficacy of the methods, techniques, and settings already known.
Multilayered Approach to Evaluate Mobile User Interfaces

to them from previous experiences. Thus, the major question that is addressed is whether it is possible to adapt the methods, techniques, and settings from previous evaluation experiences to this new class of devices, given their distinctive features.

The most frequent question raised in the vast majority of studies presented in the literature is whether to adopt a field approach or a laboratory approach. However, little is discussed in terms of which techniques are best suited for the specific evaluation target and its context of use. While this polemic subject may represent to the HCI specialist an important concern, it is equally important to consider the efficacy of the method, which accompanies this choice of approach (efficacy meaning the quality of the answers to the questions formulated as the basis of the evaluation). This is because the efforts employed in the evaluation may not pay off if a method is not well chosen or well employed.

This chapter presents a method for evaluating mobile devices based upon a set of techniques already known to the HCI specialist community. Each technique evaluates the problem from different perspectives: the user perspective (expressed as views on the product obtained through a questionnaire), the specialist’s perspective (expressed when analyzing the user performance during the usability evaluation), and the usability community perspective (expressed in the form of standards conformity assessment). Each of these perspectives identifies evaluation problems and, when overlaid, they lead to a more reliable and complete product appraisal.

The remainder of this chapter is structured as follows. The second section gives a brief overview of the evaluation approaches currently in use for mobile devices, according to the literature review. The third section outlines the multi-layered approach. The fourth section illustrates the application of the multi-layered approach by means of a case study involving a Personal Digital Assistant (PDA). The fifth section discusses the results of the case study and their implications for the questions posed in this chapter. Finally, the sixth section concludes with the discussion of future trends in evaluation methods and how to apply the existing experience to the evaluation of this new class of products.

USER INTERFACE EVALUATION FOR MOBILE DEVICES

In the context of user-centered design processes, a significant portion of usability work involves the coordinated acquisition of valid and reliable data by a team of professionals. These specialists have varied backgrounds and skills and employ a number of evaluation methods. The expected result is an improved system design. This is achieved by the successful identification of a system’s usability problems that might impact the interaction quality for a range of users.

Usability data consists of any information that can be used to measure or identify factors affecting the usability of a system being evaluated (Hilbert & Redmiles, 2000). These data are crucial for designing successful systems intended for human use. Such data are gathered by usability evaluation methods and techniques that can assign values to usability dimensions (Rosson & Carroll, 2002) and/or indicate usability deficiencies in a system (Hartson, Andre, & Williges, 2003). According to the International Organization for Standardization (ISO, 1998), usability dimensions are commonly taken to include user efficiency, effectiveness, and subjective satisfaction with a system in performing a specified task in a specified context.

Usability data are gathered via either analytic or empirical methods (Nielsen, 1993; Mayhew, 1999; Rosson & Carroll, 2002). Analytic methods, in which a system is evaluated based on its interface design attributes, are usually conducted by HCI specialists and do not involve human participants performing tasks. This means that these
Multilayered Approach to Evaluate Mobile User Interfaces

methods often rely on the specialists’ judgment. Empirical methods, in which the system is evaluated based on observed performance in actual use, involve data collection of human usage.

Other classifications include direct methods (recording actual usage) and indirect methods (recording accounts of usage) (Holzinger, 2005). There are also formative and summative methods (Wixon & Wilson, 1997). The direct methods are used to generate new ideas and gather data during the development of a system in order to guide iterative design (Hix & Hartson, 1993). The indirect methods are used to evaluate existing systems and gather data to evaluate a completed system in use (Scriven, 1967). Discovery methods (also called qualitative methods) are used to discover how users work, behave, and think, and what problems they have. Decision methods (also called quantitative methods) are used in selecting a design among several alternatives or in picking elements of interface designs (Wixon & Wilson, 1997).

In essence, usability data have been classed in a number of other models and frameworks, often focusing on (1) the approach employed for gathering the data (including the resources expended and the degree of formality) (Danielson, 2006); (2) the context of use (including lighting, noise level, network connectivity, communication costs, communication bandwidth, and the social situation) (ISO, 1998; ISO, 1999; Jones & Marsden, 2006); (3) the nature and fidelity of the artifact being evaluated (EATMP, 2000); and (iv) the goal of the acquisition process (Kan, 2002).

It is a fact that usability evaluation for stationary computer systems has grown in the last two decades. In spite of debates still taking place within the HCI area, they are often based on a tacit understanding of basic concepts. One example of this understanding is in relation to the distinction between field and laboratory evaluation approaches and their importance to the area. Classical extensive guidelines were written that describe how usability evaluation in controlled environments should be conducted (e.g., Dumas & Reddish, 1999; Mayhew, 1999; Nielsen, 1993). Additionally, experimental evaluations of the relative strengths and weaknesses of different techniques are available that can be applied in a usability evaluation (e.g., Molich et al., 1998).

In the last decade, methodologies and approaches in HCI have been challenged by the increasing focus on systems for wearable, handheld, and mobile computing devices. One such move beyond office, home, and other stationary-use settings has pointed to the need for new approaches in designing and evaluating these systems (Kjeldskov, 2003). While the primarily task-centered evaluation approaches may be applicable to the desktop computing paradigm (often structured with relatively predictable tasks), they may not be directly applicable to the often-unpredictable continuous interaction possibilities and relatively unstable mobile settings. Additionally, it is not easy for evaluation methods to integrate completely or even adequately in real world or simulated settings contexts during the evaluation process. Authors argue that mobile computing demands not only real users but also a real or simulated context with device interaction tasks. It also demands real tasks or realistic task simulations.

There are a number of studies that discuss the question of whether the evaluation should be carried out in a laboratory or field context (e.g., Goodman et al., 2004; Kjeldskov & Stage, 2004; Kjeldskov et al., 2005; Po et al., 2004). All of these papers have a common theme, in that they apply a multi-method approach to performance measurement and discuss solutions for efficient data analysis. Nonetheless, it is important to note that the approach to usability evaluation depends on the relevance of the results presented as well as on the quality of the data analysis process. In general, the reports only present the results of the data analysis, omitting the details of the analysis process itself. While the data gathering method is critical for data quality, a more rigorous analysis
on user comments and problem reports could help specialists better assess their choices.

There is a lot of current human-computer interaction research on alternatives for data collection methods and techniques. However, adequate data analysis and validation are only presented in few cases (e.g., Nielsen, 1994; Dumas & Redish, 1999; Po et al., 2004). In general, this aspect of the HCI research is poorly described in the literature, there being only vague conclusions and little guidance for attempts at successfully replicating the findings in other evaluation contexts. Many methods and techniques have been employed in the analysis of empirical data gathered during usability evaluations. Examples are for field testing analysis, video data analysis (Sanderson & Fisher, 1994), expert analysis (Molich et al., 1998), and head-mounted video and cued recall (Omodei et al., 2002). Its time-consuming character and its poor applicability for industrial purposes can explain the absence of an in-depth usage data analysis when under resource constraints (Bailie & Schatz, 2005). Nonetheless, it is strongly recommended for research purposes as a means to support new findings. For the same reason, it is equally important to provide sufficient detail to allow for replication and a substantiated choice of methods with similar levels of description.

THE MULTILAYERED EVALUATION APPROACH

The method described here was originally proposed for evaluating desktop interfaces. It was then adapted to evaluate the usability of mobile devices. It is based upon a multi-layered approach that combines standard conformity assessment, user performance measurement, and user satisfaction measurement. Each one of these evaluation techniques detects problems from a specific point of view. The multilayered approach is based on the premise that the combination of techniques (triangulation) will produce complementary and more robust results.

Standard Conformity Assessment

According to the International Organization for Standardization (ISO), conformity assessment means checking whether products, services, materials, processes, systems, and personnel measure up to the requirements of standards (ISO, 2006).

In its original version, this evaluation method adopts the standard ISO 9241 (Ergonomic Requirements for Office Work with Visual Display Terminals).

In the PDA case study it was found that only some parts of this standard can be applied to this mobile device: Parts 11 (ISO 9241-11, 1998), 14 (ISO 9241-14, 1997), 16 (ISO 9241-16, 1999), and 17 (ISO 9241-17, 1998). There are also some other standards that apply to this kind of device such as the ISO/IEC 14754 (Pen-based Interfaces—Common gestures for text editing with pen-based systems) (ISO/IEC 14754, 1999) and others that, although applicable to mobile devices, do not apply in this specific case. Examples are the ISO/IEC 18021 (User interfaces for mobile tools for management of database communications in a client-server model), since it is for devices capable of performing data interchange with servers (ISO/IEC 18021, 2002); and ITU-T E.161 (Arrangement of digits, letters, and symbols on telephones and other devices that can be used for gaining access to a telephone network, also known as ANSI T1.703-1995/1999, and ISO/IEC 9995-8:1994) (ITU, 2001).

User Satisfaction Measurement

User satisfaction has received considerable attention from researchers since the 1980s as an important surrogate measure of information systems success (Aladwani & Palvia, 2002; Goodhue &
Thompson, 1995; Bailey & Pearson, 1983). While most user satisfaction measuring instruments were not Web-based at the time of development, others have been successfully validated in a Web-based environment (e.g., De Oliveira et al., 2005).

The user satisfaction diagnosis provides an insight into the level of user satisfaction with the product, highlighting the importance of the problems found and their impact on the product acceptance.

**User Performance Measurement**

The user performance measurement aims in general to provide data on the effectiveness and efficiency of a user’s interaction with a product. It enables comparisons with similar products, or with previous versions of the same product along its development. Additionally, it can highlight areas where a product can be enhanced to improve usability. When used with the other methods, the evaluator can build a complete picture of the usability of a system.

The most significant user interface problems can be found by conducting experiments (usability tests) with representative users to observe how quickly, easily, and safely they can operate a product. The major change introduced in the original method concerns the introduction of field tests as a complement to the original laboratory tests.

**The Experiment: Comparing Field and Laboratory Use of a PDA**

The main objective of this study is to investigate the need for adapting the original evaluation method to the context of mobile devices, based on the analysis of the influence of the context (field versus laboratory and mobility versus stationary interaction) on the evaluation of mobile devices and applications.

The mobile device chosen as the target for this study was a PDA, the *Nokia 770 Internet Tablet* and some of its native applications. Tests were performed in a controlled environment (the usability laboratory) and also in the field. Twenty-four users took part in the experiment, divided into two groups of twelve.

**Experiment Design**

The study was designed to investigate the influence of the context (field and laboratory) and associated aspects such as mobility, settings, and so forth, and the user experience on the evaluation results. The independent variables are those that are not influenced by the context, by the test facilitator, or by external factors such as noise and lighting. An experiment plan was drawn from the study’s objectives. The independent variables were chosen as follows:

- **Task context** comprises factors that may affect the users’ behavior and their performance during the experiment (usability test). These factors may be internal or external to the user. The external factors originate in the field environment, examples being noise level and light intensity. The internal factors, on the other hand, are stress or other health conditions that may affect the user’s mental and physical abilities.
- **User mobility** refers to the conditions under which the task is being performed. An example is if the user is required to work while being mobile, that is, moving between places or wandering while working.
- **User experience level** refers to the user’s knowledge regarding mobile devices in particular and desktop computers systems in general.

The dependent variables are all dependant on the user’s experience level:
Multilayered Approach to Evaluate Mobile User Interfaces

- **Task time** represents the time taken by a device’s user to perform a task.
- **Number of incorrect choices** measures how many times the user has made incorrect choices while selecting options in the interface through a menu dialogue.
- **Number of incorrect actions** measures how many times the same error (excluding the number of incorrect choices) was committed by the user while performing a task.
- **Number of accesses to the online help** and **number of accesses to the printed help** measure how many times the user accessed the online and printed help while performing a task.
- **Perceived usefulness** represents the user’s opinion about the usefulness of the mobile application for the prescribed task.
- **Perceived ease of use** represents the user’s subjective satisfaction when using the mobile device.

Table 1 summarizes the experiment plan, which states the independent and dependent variables to be observed during the experiment and used as indicators to answer the research questions.

**Test Environment**

A software tool was used in the field environment to remotely capture the device’s screen through a wireless connection to the lab. The user inputs (through keypad and stylus) were registered by a micro-camera coupled to the device and also remotely connected to the laboratory through a wireless connection. The interaction was regis-

---

**Table 1. Plan for the experiment with the device Nokia 770**

<table>
<thead>
<tr>
<th>EXPERIMENT PLAN</th>
</tr>
</thead>
</table>
| **Target-Problems** | 1. With the shape/dimensions of the product  
2. With the mechanisms for information input/output  
3. With the processing power  
4. With the navigation between functions  
5. With information legibility |
| **Test Objectives** | 1. Investigating the target problems  
2. Detecting other problems |
| **Objective Indicators** | 1. Task execution time  
2. Number of incorrect actions  
3. Number of incorrect choices  
4. Number of repeated errors  
5. Number of accesses to the online help  
6. Number of off-line help (printed manuals) accesses |
| **Subjective Indicators** | 1. Product ease of use  
2. Task completion easiness  
3. Input mechanism ease of use  
4. Text input modes ease of use  
5. Understandability of terms and labels  
6. Understandability of messages  
7. Help mechanism efficiency |
Multilayered Approach to Evaluate Mobile User Interfaces

tered in the controlled environment using two
video cameras installed in the laboratory. One
was focused on the users’ facial expressions and
the other registered the device screen. As in the
field environment, software was used to remotely
capture the device’s screen. Since the field setting
required a wireless network, the field experiment
was performed in the area surrounding the univer-
sity’s computer department. In both cases, the test
facilitator was a human interface specialist who
remained within reach in case the user required
any explanation on the test procedure.

Participants

Users participating in the PDA experiment were
selected on the basis of having previous experi-
ence with mobile devices (such as mobile phones),
computers, and the Internet. They were also re-
quired to have some familiarity with the English
language, since this is the language adopted in
the device’s user interface and in its documenta-
tion. The user sample was then classed according
to the users’ experience level into the categories
shown in Table 2.

The recruited users were divided into two
groups of 12 to participate in the field and labora-
tory tests. Based on user experience level, both
groups were then subdivided into three subgroups
of four beginners, four intermediates and four
experts.

Materials

Laboratory Test Materials

• **Hardware**: The Nokia 770 Internet Tablet;
  PC based Workstation (2); Video cameras
  (3); Microphones (2).

• **Software**: VNC (Virtual Network Com-
  putting) software to capture the screens
during the interaction with the device; the
WebQuest tool with the questionnaires pre-
test (to gather the user profile) and post-test
(to collect and process the user satisfaction
level).

• **Miscellaneous**: The Nokia 770 Internet
  Tablet Manual; chronometer (1); CDs for
  video backup; participant registration form;
test conditions acceptance forms on which
  the users declared their acceptance of the
  experiment conditions; task script that con-
sists of a written task description to guide
  the user during the session (versions for the
  user and for the evaluator); Form for event
  log.

Field Test Materials

• **Hardware**: The Nokia 770 Internet Tablet;
  PC-based Portable (laptop) Workstation (1);
  wireless video micro-camera (1); apparatus

Table 2. User sample categorization

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>Beginner</th>
<th>Intermediate/Advanced</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Computer Knowledge</td>
<td>Basic/Intermediate</td>
<td>Intermediate/Advanced</td>
<td>Intermediate/Advanced</td>
</tr>
<tr>
<td>Previous Experience with Nokia 770</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
to support the video micro-camera (1); television set (1); VCR equipment (1).

- **Software:** VNC (Virtual Network Computing) software to capture the screens during the interaction with the device; WebQuest tool with the questionnaires pre-test (to gather the user profile) and post-test (to collect and process the user satisfaction level).

- **Miscellaneous:** Chronometer (1); CDs and VHS tapes for video backup; participant registration form; test conditions acceptance forms on which the users declared to accept the experiment conditions; task script that consists of a written task description to guide the user during the session (versions for the user and for the evaluator); form for event log.

**Camera Apparatus**

The apparatus shown in Figure 1 was built to couple a video micro-camera to the mobile device. This allowed the recording of user interaction through a remote link with the laboratory computer.

**The WebQuest Tool**

A Web tool named *WebQuest* supports the method application (De Oliveira et al., 2005). This tool was developed to support the specialist during data collection, to provide automatic score computation, to perform statistical analysis, and to generate graphical results. *WebQuest* also enables the specialist to reach a more diverse and geographically widespread sample of users through the Internet. One of its features is a flexible questionnaire structure, which enables specific context adaptation and, by means of an estimation model, ensures a higher degree of confidence on the indicators of user satisfaction. Currently *WebQuest* supports two questionnaires: (i) a pre-test questionnaire, the *USer* (*User Sketcher*), conceived to raise the profile of the system users; and (ii) a post-test questionnaire, the *USE* (*User Satisfaction Enquirer*), conceived to raise the user degree of satisfaction with the system. The pre-test questionnaire incorporates a model to estimate the user’s subjective satisfaction and can be answered directly on the Web. The questions are related to the users’ physical characteristics, knowledge, and skills. Both questions and answers are configurable.

As for the *USE* (*User Satisfaction Enquirer*), it allows gathering quantifiable variables on the user acceptance of the device. Three of its aspects are of special interest. Firstly, it incorporates a model to estimate user subjective satisfaction. Secondly, the questionnaires can be answered directly on the Web. Thirdly, the items are partially or totally configurable. The adoption of an estimation model by *USE* allowed us to establish a subjective satisfaction coefficient directly from the inspection of the respondents’ samples. The *WebQuest* tool allows the specialist to easily edit the questionnaire’s items. These items are orga-
nized into groups: (1) fixed, which are applicable to various evaluation contexts and thus are not allowed to be altered; (2) semi configurable, which allow for changes in the answer options; and (3) configurable, which can be fully configured (both the question and respective options of answers). USE supports the specialist from the data collection through to automatic score computation, performing statistical analysis, and generating graphics with the results.

**Experiment Procedure**

The techniques employed in the experiment procedure were the observation and subsequent video analysis for accumulating quantitative data (such as time spent and error rate). An automated video capturing tool recorded the interactions of the subjects during the field tests to ensure a non-intrusive observation method. During task execution, the users were asked for their consent before being filmed. The conditions of test-subject participation included a written commitment not to disclose any product information. The users were also asked to give consent so that their images and/or sound recordings made during the experiment could be used for research purposes or in a multimedia product evaluation report. On the other hand, the users were given assurances from the evaluation team that no personal information or individual performance results would be disclosed.

The first step in following the method consisted in defining the evaluation scope for the product as well as a scenario for the test. Table 3 illustrates the sequence of tasks performed during the experiment.

The decision was based on a heuristic evaluation performed by the evaluation team. This initial step also supports the definition of a general profile for the user sample and a classification into categories. Following, the method the evaluation objectives were defined. These became the basis for choosing the product evaluation scenario (product context of use and laboratory settings) and the corresponding tasks to be performed by the users during the experiment. Having planned the evaluation, a pilot test was conducted to verify the adequacy of the proposed experiment procedural, materials, and environment. Through this fine tuning procedure it was found, in the PDA case study, that the time to perform the tasks had been underestimated. This resulted in re-dimensioning the test scenario to six tasks, with a review of the tasks themselves to fit the established session time of sixty minutes to prevent user tiredness.

All subjects were submitted to the same procedure prescribed in the experiment protocol. The study was conducted first in a laboratory setting and then in the field environment. During the field tests the participants were taken outdoors, and the tasks were conducted in an environment that was as close to real-use conditions as possible.

The experiment conducted in the usability laboratory had the audio and video of each session recorded. In the field experiment, only the video of the sessions was recorded, supplemented by comments written by the specialist. As described in the experiment protocol, each session consisted of the following steps: (1) introducing

---

**Table 3. Test scenario and sequence of tasks to be performed during experiment**

<table>
<thead>
<tr>
<th>TASKS IN SCRIPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01 Initializing the device</td>
</tr>
<tr>
<td>T02 Searching for books in an online store</td>
</tr>
<tr>
<td>T03 Visualizing a PDF file</td>
</tr>
<tr>
<td>T04 Entering textual information</td>
</tr>
<tr>
<td>T05 Using the electronic mail</td>
</tr>
<tr>
<td>T06 Using the audio player</td>
</tr>
</tbody>
</table>
the user to the test environment by explaining the test purpose, the procedure to be followed and the ethics involved in terms of the conditions of participation; (2) applying the pre-test questionnaire; (3) performing the task script; (4) applying the post-test questionnaire; and (5) performing a non-structured interview.

At the time of the experiment, the Nokia 770 Internet Tablet device was not yet widely known in the Brazilian market. The users who claimed to have had no previous contact with it were given a quick introduction. This introduction consisted of an instructional material given to the recruited users and also a quick explanation about the device’s input and output modes and its main resources.

Results

The results obtained from the experiment in which the multi-layered method was applied support the original assumption that, in spite of the distinctive features of this class of devices, it is possible to adapt from the evaluation experience with conventional devices. This conclusion is supported by the evidence that the evaluation context did not significantly influence the user performance or the opinion about the device’s usability, given through the analysis of the objective and subjective indicators associated with the experiment.

Standard Conformity Assessment Results

The results of the conformity assessment to the standards ISO 9241 Parts 14 and 16 and ISO 14754 are illustrated in Table 4. According to ISO, conformity assessment results can be summarized by computing an adherence rate (AR). This is the percentage of the applicable recommendations (Ar) that were successfully adhered to (Sar).

In spite of the device’s characteristics that limit the number of applicable recommendations, these results corroborate the idea that the standards inspection is still applicable in the evaluation process. The efficacy of this technique can be considerably improved if it is based upon standards conceived specifically for mobile devices, which could evidence more usability problems.

User Satisfaction Measurement Results

For the PDA case study context, both questions and answers of the USE questionnaire were configured. The questionnaire was applied soon after the usability test and answered using the mobile

Table 4. Nokia 770 conformity assessment with standards

<table>
<thead>
<tr>
<th>Standard</th>
<th>#Sar</th>
<th>#Ar</th>
<th>AR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO 9241 Part 14</td>
<td>45.0</td>
<td>53.0</td>
<td>84.9</td>
</tr>
<tr>
<td>ISO 9241 Part 16</td>
<td>26.0</td>
<td>33.0</td>
<td>78.8</td>
</tr>
<tr>
<td>ISO 14754</td>
<td>4.0</td>
<td>11.0</td>
<td>36.4</td>
</tr>
</tbody>
</table>

Sar—Successfully adhered recommendations
Ar—Applicable recommendations
AR—Adherence Rate

\[ AR = \frac{Sar}{Ar} \times 100\% \]
device itself. As mentioned before, its purpose was to collect information on the user’s degree of satisfaction with the device and on aspects such as interface navigation, documentation, and overall impressions.

The USE was composed of three sections. The first section is relative to “the product Use and Navigation.” It is composed of 17 items and focuses on aspects such as menu items, navigation between functions, understandability of the messages, ease of use of the basic functionalities, and of the device’s input and output mechanisms. The second section consists of six questions related to the online and off-line (printed manuals) documentation. The last section (“You and the product”) consists of 15 items and aims to get the user’s impressions and product acceptance level. The first 23 items use a 5-point semantic scale (1: very easy; 2: easy; 3: not easy nor difficult; 4: difficult; and 5: very difficult). The last 15 items use another 5-point semantic scale (1: completely agree; 2: agree; 3: do not agree nor disagree; 4: disagree; and 5: completely disagree). The users were asked to answer the questions and to assign an importance level to each one of them, on a scale from 0 to 10.

For the post-test questionnaire, USE adopts the model proposed by Bailey and Pearson (Bailey & Pearson, 1983) for measuring the overall user’s sense of satisfaction. The following adaptations to the dimensions were considered: (1) the association of only one (1) semantic differential scale to the items, instead of the four (4) semantic differential scales, as proposed in the original model; (2) the adoption of a 5-point Likert scale, delimited by the ends -2 and 2 (instead of the 7-point scales delimited by the ends -3 and 3 as originally proposed); and (3) the incorporation of a 11-point importance scale (0 corresponding to non applicable), varying from 0.0 to 1.0 in intervals of 0.1 (instead of the original 7-point scales, which varied from 0.1 to 1.0 in intervals of 0.15).

The user’s subjective satisfaction indicators for the PDA case study were 0.330 for the laboratory experiment and 0.237 for the field experiment. The normalized value ranges of the user satisfaction concerning a product are 0.67 to 1.00 (Extremely Satisfied), 0.33 to 0.66 (Very satisfied), 0.01 to 0.32 (Fairly satisfied), 0.00 (Neither satisfied nor unsatisfied), 0.01 to 0.32 (Fairly dissatisfied), 0.33 to 0.66 (Very dissatisfied), and 0.67 to 1.00 (Extremely dissatisfied). This is in accordance with the Bailey and Pearson model (Bailey & Pearson, 1983). The results obtained correspond respectively to Very satisfied and Fairly satisfied.

**Performance Measurement Results**

**The User Sample Profile**

The user sample profile was drawn with the support of the questionnaire USer. It was composed of 13 male and 11 female users, of which eight were undergraduate students, 12 post-graduate students, two graduate level, and two post-graduate level. The ages varied between 18 and 29 years. They were mainly right handed and mostly used some sort of reading aid (either glasses or contact lenses). All of them had at least one year of previous experience of computer systems and were currently using computers on a daily basis.

**User Performance Data Analysis**

After having analyzed the data gathered during the experiment on the user performance and having analyzed the list of problems found with this technique, it was possible to evaluate their impact and class them as: minor (50%), medium (50%), major (0%), consistency (35.7%), recurrent (64.3%), and general (0%).

The data analysis consisted of a statistical processing and had two main purposes: (1) to investigate the influence of the context on the results of the evaluation method (through the comparison of the results obtained from both environments); and (2) to investigate the influence of the user experience with the mobile device on
the test results within each context. For the latter purpose, the three categories illustrated in Table 2 were used.

The statistic analysis performed consisted of: (1) building a report with univariance statistics; (2) generating the covariance matrices for the objective and subjective indicators that were previously defined; (3) applying the one-way F ANOVA test (Tabachnick & Fidell, 2006) to the data obtained from the previous step in order to investigate possible differences; and (4) applying the Tukey-Kramer process (Tabachnick & Fidell, 2006) to the one-way F ANOVA results aiming to investigate if the found differences were statistically significant to support inferences from the selected sample. The result of this technique was the identification of 13 problems, of which 92.3% were found in the laboratory and 61.5% in the field as: Laboratory (38.5%); Field (7.7%); and Laboratory & Field (53.8%).

Overlaying Results

Since the multi-layered evaluation is based upon a triangulation of results, Table 5 summarizes the usability problem categories identified by the three techniques.

The numbers correspond to the identification of each problem from a list of problems found through each technique. As can be seen from Table 5, some of the usability problem categories were more related to the performance measurement (e.g., hardware aspects, help mechanisms, processing capacity) whereas others (e.g., menu navigation, presentation of menu options) were identified by the conformity assessment. It was possible to identify 66.7% of the problems found by other methods when combining the results from the post-test questionnaire with the user comments made during the experiment and the informal interview at the end of the experiment. This confirms the importance of combining techniques to obtain a more complete result when performing usability evaluation. It must be pointed out that 29.62% of the problems based on the user opinion about the product were in disagreement with the results of the other two evaluation dimensions (specialist and the community points of view). This discrepancy can originate from the users’ perception of product quality and the perception of their own skills to perform the task, accepting full responsibility over the difficulties that might arise during the interaction. When overlaying the problems in Table 5, in the category Menu navigation, the same problem was found by the techniques Standards Inspection and Performance Measurement.

DISCUSSION

From this study’s data analysis it became evident that certain problem categories are better found by specific techniques, as shown in Table 5. For instance, problems associated to the device’s physical characteristics are better found by means of conformity assessment, whereas the user performance located problems associated to the device’s applications.

The analysis of the pre-test and post-test questionnaires and the informal interviews showed that domain knowledge and computer literacy have significant influence on user performance with mobile devices. This was true both under laboratory conditions and in the field, in relation to the incidence of errors. The univariate analyses of variance of the performance variables: Time, Errors, and Accesses to help, are presented in Table 6.

From this table, it can be seen that the user experience level had a more significant effect on the number of errors in the field experiment than in laboratory experiment.

The studies in the literature fit basically into two categories: (1) user mobility, which means moving while using the device (inside of a laboratory or outdoors) and (2) user attention division. However, this study considers both aspects as...
Table 5. Overlay of results obtained with the three evaluation techniques

<table>
<thead>
<tr>
<th>PROBLEM CATEGORY</th>
<th>SI</th>
<th>PM</th>
<th>SM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location and sequence of menu options</td>
<td>✓ (05)</td>
<td></td>
<td>× (05)</td>
</tr>
<tr>
<td>Menu navigation</td>
<td>✓ (02)</td>
<td>✓ (01)</td>
<td></td>
</tr>
<tr>
<td>Presentation of menu options</td>
<td>✓ (02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information feedback</td>
<td>✓ (01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Object manipulation</td>
<td>✓ (05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symbols and icons</td>
<td>✓ (02)</td>
<td>× (02)</td>
<td></td>
</tr>
<tr>
<td>Text entry via stylus (Writing recognition)</td>
<td>✓ (07)</td>
<td>✓ (01)</td>
<td>✓ (08)</td>
</tr>
<tr>
<td>Text entry via virtual keyboard</td>
<td>✓ (01)</td>
<td>✓ (01)</td>
<td></td>
</tr>
<tr>
<td>Processing power</td>
<td>✓ (02)</td>
<td>✓ (02)</td>
<td></td>
</tr>
<tr>
<td>Hardware issues</td>
<td>✓ (03)</td>
<td>✓ (03)</td>
<td></td>
</tr>
<tr>
<td>Fluent tasks execution</td>
<td>✓ (05)</td>
<td>✓ (05)</td>
<td></td>
</tr>
<tr>
<td>Online and offline help</td>
<td>✓ (01)</td>
<td>✓ (01)</td>
<td></td>
</tr>
</tbody>
</table>

Legend:
SI—Standards Inspection
PM—Performance Measurement
SM—Satisfaction Measurement
✓ - Consistent findings
× - Contradictory findings

Table 6. Influence of the user experience on the performance indicators: Time, Number of errors, and accesses to help

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Dependent variable</th>
<th>p-Value (Lab)</th>
<th>p-Value (Field)</th>
<th>Significance (α=0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>Task Time</td>
<td>0.081</td>
<td>0.081</td>
<td>Not significant</td>
</tr>
<tr>
<td>Experience</td>
<td>Errors</td>
<td>0.011</td>
<td>0.002</td>
<td>Significant</td>
</tr>
<tr>
<td>Experience</td>
<td>Help Accesses</td>
<td>0.427</td>
<td>-</td>
<td>Not significant</td>
</tr>
</tbody>
</table>
part of the task context. In this experiment, the field test subjects were free to choose between moving or remaining still as they performed the task with the mobile device. During the informal interview the users stated that in a real context they would not perform the experiment tasks on the move, since they demanded too much attention. The specialist encouraged users to wander around the environment, although they could choose to enter a room in the building, sit down, or even lay the device on a table (which they did in most cases, under the argument that this setting was more comfortable). The movement registered was limited to situations in which the user waited for some device processing. (e.g., Web page downloads). There was a clear interference of the environment on the user attention during the field tests while moving.

The device’s physical characteristics affected the user performance and the data gathering during the experiment. Outdoors, in ambient light, the device’s legibility was reduced and aggravated by the reflections on the screen. According to the user’s opinion stated during the informal interview, the camera apparatus did not interfere with the task execution, but the majority decided to lay the device down during task execution.

As for the entry of text information, the users showed a preference for the virtual keyboard instead of hand written character recognition. Based on their comments, as well as on the informal interview, it was concluded that writing long messages is very cumbersome both using the virtual keyboard and using the handwriting recognition application. Confirming previous findings, the experiment demonstrated that applications that require a lot of interaction and user attention are inappropriate for performing while walking due to attention division. This conclusion reinforces that, for the device targeted in this study, in spite of its mobility, the evaluation settings did not need to differ substantially from the one employed in the evaluation of stationary devices since the users tend not to wander while performing tasks that demand their attention or consisted of text input.

Until recently, studies have been published which deal with new paradigms and evaluation techniques for mobile devices. Few of the proposed new techniques are really innovative if compared to the ones traditionally employed. On the other hand, the main argument for proposing new techniques concerns the user and device mobility and the influence of this mobility on user performance. In contrast, this study evaluated the effect of mobility not only from the user performance perspective but also from user opinion point of view and the user level of satisfaction. From the application of the multi-layered approach, the data gathered and analyzed support the initial assumption that minor adaptations in the traditional evaluation techniques and respective settings are adequate to accommodate the evaluation of the category of mobile devices targeted by this study.

The conclusions corroborate with the views of the authors and that of Po (Po, 2003) that the laboratory and field evaluations do not diverge but are complimentary. As shown in this study, they both add to the evaluation process, producing data that is significant to the process and reinforcing the relevance of a multi-layered approach for the usability evaluation of mobile devices.

FUTURE TRENDS

Mobile devices impose challenges to the usability evaluation that are unique in respect to the observation strategies and the conception of test scenarios. With the continuous technological advances, a wider variety of new devices is being released into the market, challenging users with the complexity of the interaction. In this scenario, the importance of the product usability is undisputable as is also the correct choice of evaluation methods, techniques, and tools.

One emerging trend in the mobile devices evaluation field is the possibility of gathering data
in an unobtrusive way, using tools for remote, and automatic data capture that are transparent to the user. Developing those tools is a challenging activity given the inherent restrictions presented by the mobile devices (such as their limited processing power and limited storage capacity). But, in spite of the current limitations, it was shown in this study that the tools are becoming available to provide a great contribution to the evaluation setup and that these tools would benefit from further development.

REFERENCES


Organization for Standardization, Geneva, Switzerland.


ITU-T E.161. (2001). *Arrangement of digits, letters and symbols on telephones and other devices that can be used for gaining access to a telephone network*. International Telecommunications Union-telecommunications, Geneva, Switzerland.


Multilayered Approach to Evaluate Mobile User Interfaces

laboratory and field testing. *Human-Computer Interaction, 9*, 251–317.


**KEY TERMS**

**Conformity Assessment:** A collective term used for a number of techniques used to determine if a product, system, or process (including design) meets a defined specification.

**Device Mobility during a Usability Evaluation:** The ability to interact with the user and continue to perform its functions while being transported.

**Efficacy of an Evaluation Method or Technique:** Translated into the number of problems found, gravity of those problems versus the time, and cost of performing the experiments.

**Likert Scale:** An attitude scale in which respondents indicate their degree of agreement/disagreement with a given proposition concerning some object, aspect, person, or situation.

**Multi-Layered Evaluation Approach:** A product or prototype usability evaluation method that combines techniques for data gathering and analysis based on multiple perspectives (the user’s, the specialist’s, and the usability community). The results are overlaid in order to find discrepancies and offer more robust results.

**User Mobility during the Usability Evaluation:** The ability to move while performing a task with a product.

**User Performance Measurement:** The process of gathering actual data from users as they work with a system and its documentation. Usually, the user is given a set of tasks to complete and the evaluator measures the relevant parameters such as the percentage of tasks or subtasks successfully completed, time required to perform each task or subtask, frequency and type of errors, duration of pauses, indications of user frustration, and the ways in which the user seeks assistance.

**User Satisfaction Measurement:** The process of obtaining qualitative and quantitative information which indicates the extent to which user expectations concerning some object, process, product, or situation are being met. Such information can be obtained in a variety of ways, both formally and informally.

**Virtual Network Computing (VNC):** A desktop sharing system that uses the RFB (Remote Frame Buffer) protocol to remotely control another computer. It transmits the keyboard presses and mouse clicks from one computer to another over a network, relaying the screen updates back in the other direction.

This work was previously published in the *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, edited by J. Lumsden, pp. 847-862, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 8.5

Digital Imaging Trek: A Practical Model for Managing the Demand of the Digitally Enabled Traveller

Stephen C. Andrade
Johnson & Wales University, USA

Hilary Mason
Johnson & Wales University, USA

ABSTRACT

This chapter introduces the concept and activities of the digitally enabled tourist and the impact such as tourist has on the travel and tourism industry. It summarizes the existing and emerging technical environments that encourage the use of hand held digital recording devices and personal Internet communications. Additionally, it looks at ways tourists publish and exhibit digital visual and written artifacts of their travel experience. The chapter introduces general types of digital communication infrastructure to be considered by the industry to create an experience to support this type of tourism. The authors hope that further understanding of the digitally enabled tourist will inform travel professionals to better facilitate commerce and practice in the industry.

INTRODUCTION

Technology has had a great influence on the tourism and travel industry. In recent years digital communication platforms and technologies have evolved and become accessible to a wide mainstream audience of tourists and travelers. Organizations engaging in travel and tourism commerce need to understand this wave of innovative behavior among their customers. From cell phones to high speed wireless Internet connection to the variety of creative ways customers have applied digital technologies—the travel and tourism industry is being widely affected by these emerging trends in information and communications technologies (ICT). Is the digitally enabled traveler the wave of the future? How will companies and services shift business models to optimize the experience of people with digital
devices? Are there commercial opportunities embedded in these? What does a traveler need to know to keep current with changing technologies? It is critical that services providers and travelers alike stay informed, because one thing is certain, technological innovation and change will be a constant companion for the travel and tourism industry.

This chapter provides insight into technology trends that will be helpful to the practitioner, student, educator and the tourist-travelers themselves. Being prepared to meet the new demands of customers will provide rewarding experiences for parties on all sides of the tourism equation.

BACKGROUND

Since the wide spread use of the telephone in the 1920s, information and communications technologies (ICT) have had a great influence on the industry of tourism. In the 1990s, the wide spread use of powerful desktop computers, enterprise wide systems, and the World Wide Web (WWW) continued to transform the way business was conducted in all facets of the travel and tourism. While ICTs have had a dramatic impact on the mechanics of tourism business practices, the virtual explosion of new inexpensive digital communication technologies is transforming the experience of tourism from the traveler’s point of view.

The new generation of hand held mobile technologies, the expansion of wireless (WiFi) networks and the surge in digitally hosted social interchange services present new opportunities for engaging all sectors of the tourism industry. This new generation of technologies also presents new challenges to the industry to structure services with the digital service users in mind.

New services that provide online collaborative and social interaction through the World Wide Web now shape and influence vast communities of millions of online customers. Traveler-centered mobile technologies are increasingly used for many activities embedded in the tourist and traveler experience. This chapter will explore the model of the digital imaging trek and the digitally enabled traveler as a way to structure experiences to satisfy the demand of the technically savvy traveler in a world of advancing mobile technologies and online services. It also will provide basic technical background about the devices and infrastructure that drives these technological innovations.

Understanding the use of advanced hand held devices from the traveler’s point of view is becoming more critical to tourism industry providers. Both new and veteran professionals in the tourism industry need to think about how these technologies influence the customers’ choices, activities and ultimately their economic decisions about tourism. Customers have eagerly adopted the cheap and easy to use digital technologies. Tourism professionals need to understand the role mobile digital technologies play in the expanding global tourism field. For customers, mobile digital technologies are helping to shape the experience of tourism from initial research of a destination, through the reservation process to the final visual record of the experience.

Mobile digital technologies include a wide array of products that are inexpensive and easy to use. When mobile digital technologies are used to shape a tourist’s experience they become powerful digital communication tools reaching out to an infinite audience of like-minded users on the World Wide Web. New products in cellular telephony, visual still imaging, motion and sound recording, wireless digital connections to the World Wide Web and the services on the World Wide Web, all converge to offer a digital environment unlike any before it. The combinations of these technologies are dynamic, unique and ever expanding. These digital technologies are in a constant state of enhancement—services become faster, devices become more powerful and feature-rich, and prices come down. This
Moore’s Law is the observation made in 1965 by Gordon Moore, co-founder of Intel. In it he states that the number of transistors per square inch on integrated circuits had doubled every year since the integrated circuit was invented. Moore predicted that this trend would continue for the foreseeable future. In subsequent years, the pace slowed down a bit, but data density has doubled approximately every 18 months, and this is the current definition of Moore’s Law, which Moore himself has blessed. Most experts, including Moore himself, expect Moore’s Law to hold for at least another two decades. (http://www.webopedia.com)

Consider what it means to double the capacity of digital technologies every 12-18 months. Few if any physical systems or resources in the world of tourism have this capacity and potential. Moore’s Law not only means faster and better technology, it also enables dynamic changes in the way people use technology. In fact, it is difficult to predict exactly how consumers will respond to innovative devices and new digital gadgets. Tourism professionals need to understand this dynamic process and prepare for the digitally enabled traveler.

Who is the Digitally Enabled Traveler?

The digitally enabled traveler is a new breed of traveler equipped with devices, connectivity, skill
and motivation to create and access real-time, online, rich media knowledge bases of travel and tourism experience.

The digitally enabled traveler is motivated by the same principles as the conventional tourist. People involved in tourism are visiting locations for leisure, recreation, sight seeing, vacation and other activities. The global travel industry hosts not only tourists, but professional travelers on professional missions as well. People travel for business, cultural, scientific, educational, governmental and other kinds of activities in the world of global tourism. Most forecasts for travel of all types indicate a steady rise of 4% a year over the next decade. Even a causal observation in any busy transportation hub, such as an airport or train station will reveal how critical mobile technologies and wireless connections are to both the recreational and professional traveler alike.

People who travel with digital devices are highly motivated to stay connected to business and social networks. Minimally, most travelers today require basic voice and Internet connectivity service to maintain contact with families and tourism providers. Travelers want to stay in touch with the sources of lodging, travel bookings, reservations and other critical contacts during their trip. Some travelers thrive on constant digital connection to not only monitor progress in plans, but also stay in touch with their virtual communities.

Beyond the basic business function and family contact, travelers are using a wide array of digital devices to capture, record, edit, and exhibit the experience of tourism and travel. The days of a single film camera used in a casual manner to take pictures of highlights are over. As the tourism industry embraces near-endless global locations for travel destinations, travelers are highly motivated to capture and share their experience with digital recording and connection devices. Why? In spite of all the technical gadgetry, it is still human nature to want to discover and share a new experience. This mode of communication has been going on for generations, only now, it is played out on a global digital stage.

Digital still and video cameras along with powerful laptops and hand held devices have opened new territory for the digitally enabled traveler. These new digital technologies also offer new service-business opportunities for the tourism industry.

Today’s digitally enabled traveler is highly motivated to stay in-touch with a virtual community of people through services on the Internet. With millions of people subscribing to World Wide Web sites that host virtual communities the trend of communicating in digital interactive space has been set. Social networking through digital services is widely accepted as common practice. This segment of the technology service industry is rapidly growing as the each successive generation matures into an economic demographic that can afford the expense of travel and tourism. Millions of people are now involved in what is called “social networking.” As counter intuitive as it may seem, the way to establish human bonds is through digital technologies. Numerous Web sites have been established as sources for self-published digital image galleries, digital video galleries, audio files, user profiles, and blogs.

Social Networking: The Latest Trend

Social networking systems are an emerging technology that is beginning to have a significant influence on how people communicate. Social networking systems are Web sites that offer a collaborative or shared virtual experience, generally around a particular theme or human interest. Visitors to the site can connect with each other through shared attributes, such as interests, activities, or geographic location.

Some social networking sites, such as Facebook (http://www.facebook.com), are designed simply to facilitate the process of making social connections. Other sites, such as Delicious
An essential quality of a social networking site is that it must allow users to share information via a network of nodes (users) and connections (users with similarities). Users also can easily identify other users with similar interests.

Most social network sites utilize the community to develop an ontology via tagging. Tagging is a simple method of allowing users to attach key words to a piece of data. Users searching for those keywords can then locate a wide variety of matching media. This works on the assumption that humans will generally choose similar terms to describe similar items. Most of the time, the assumptions works well, and people network with other people's experiences.

Tags are often represented in a “tag cloud” (see Figure 2). Larger text indicates a more popular tag. Other visualizations include history, topics, and information origin. This visual model of popularity makes these sites friendly to even the most non-technical visitors.

Social networking sites are beginning to have a large impact in business. Consider the common professional activity of attending industry-specific conferences. Attendees meet for seminars, meals, and networking. Several attendees may take photos of this event. They can upload their photos to a photo networking site such as Flickr (http://www.flickr.com), and tag them with the name of the conference. Anyone searching for the conference would be able to find a complete photographic record of the event.

Social networking has even made an impact in tourism. “Where Are You Now?” (http://www.wayn.com) is a site that connects travelers, for logging trips and comparing destinations, finding travel buddies, and making friends with like-minded people. WAYN even helps with off-line networking by allowing users to send SMS (short text messages) to each other's cell phones.

Social networking presents an innovative solution to the problem of information overload. By organizing information collaboratively, useful content filters toward the people who would most like it, while useless content is dropped altogether. Finally, it is a core technology with implications that will touch all disciplines in the years to come.

A blog, or “Web log” for short, is a collection of posts around a theme or topic collected on a Web site. At the time of this writing, blog search firm Technorati (http://www.technorati.

---

**Figure 2. Sample tag cloud**

![Sample tag cloud](image)
Digital Imaging Trek

Digital Imaging Trek (http://www.digitalimagingtrek.com/) was tracking 25.4 million blogs with 1.9 billion links.

A blog is a collection of time-stamped journal entries on any possible topic. Blogs cover every subject from politics to education to technology to one particular person’s social life. Some blogs are only of interest to a few people while others have thousands of readers daily.

The Pew Internet Survey estimates that about 11%, or 50 million people, read blogs. (http://www.pewInternet.org/PPF/r/113/report_display.asp). Eighty percent of people contacted by the Business Blog Consulting Web site, a site focused on the growing use of blogs in business, believe that blogs are not a fad. Traditional media outlets (such as the BBC) have begun to add blogs to their media offerings.

The digitally enabled traveler is motivated to share experience in an immediate, visual and highly subjective manner. For a digitally enabled traveler, the personal reaction to a destination like a museum, a historic site, or a travel adventure is typically recorded as a highly personalized written blog. The digitally enabled traveler will supplement the blog with a gallery of digital images. The gallery will then be linked to a short video, compressed for Web hosting, of the activity at the site. With a reasonably fast wireless Internet connection or access to a local Internet café, this material can be posted within minutes of the experience, or in some cases, in near real time. Digitally enabled travelers with the proper digital gear and Internet connection can produce a personalized digital stream of video, images and words. Often defined as “rich” media, the mix of all these files—sound, still, motion, text—are a critical and creative connection to a virtual world for the digitally enabled traveler. For the

Table 1. Understanding the devices, practices, skills, and motivations of the digitally enabled traveler

<table>
<thead>
<tr>
<th>Device</th>
<th>Connectivity</th>
<th>Skill</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellular Phone with low resolution camera</td>
<td>Commercial wireless network, satellite connection in</td>
<td>Easy to use, entry</td>
<td>• Basic voice communications. Real time voice conversations with social and business network. Voice mail.</td>
</tr>
<tr>
<td>function</td>
<td>remote areas.</td>
<td>level skill</td>
<td>• Basic email if function available.</td>
</tr>
<tr>
<td></td>
<td>Multifunction chips available for international</td>
<td></td>
<td>• Basic organizational information - names, addresses, telephone, fax.</td>
</tr>
<tr>
<td></td>
<td>functionality.</td>
<td></td>
<td>• Basic low resolution images, very limited storage.</td>
</tr>
<tr>
<td></td>
<td>Limited WWW access if available</td>
<td></td>
<td>• Limited to real time voice based research and basic business functions such as reservations, bookings etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Limited keyboarding if necessary.</td>
</tr>
</tbody>
</table>
Table 1. Understanding the devices, practices, skills, and motivations of the digitally enabled traveler (continued)

<table>
<thead>
<tr>
<th>Device Description</th>
<th>Commercial Description</th>
<th>Entry to moderate level, requires some experience</th>
<th>Additional Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hand held multifunction PDA-personal digital assistant</td>
<td>Commercial wireless network, satellite connection in remote areas. Multifunction chips available for international functionality. Limited WWW access if available</td>
<td>• Voice function if feature available on device • Email • Moderate organization of rich information itinerary, dates, addresses, images, url’s etc. • Low resolution imaging if available, limited storage • Wireless connectivity to basic navigating on <a href="http://WWW">WWW</a>. • Basic keyboarding for research and booking business functions such as reservations, bookings, etc.</td>
<td></td>
</tr>
<tr>
<td>Battery power laptop with wireless connectivity</td>
<td>Commercial or free wireless, internet café, locally provided high speed network connection. Satellite connection in remote areas. Wireless chip functional in international standards. Access to rich sources of information on WWW.</td>
<td>Moderate to advance, able to detect wireless signal; may require some basic problem-solving. Powerful image editing software applications</td>
<td>• Voice over IP if function available • Email, online social networking, post digital images and video • Advanced organization of rich information on device’s data sources. • Edits and stores high resolution still images and digital video with sound. • Likely to use laptop for extensive research and business functions such as reservations, bookings, etc. • Uses laptop as critical extension of travel experience. Connectivity to online, social network a priority.</td>
</tr>
<tr>
<td>Digital Still Camera</td>
<td>Direct connect to laptop or other device with Firewire. Interconnect to web through computer device.</td>
<td>Easy to expert level skill</td>
<td>• Record images • Record technical file type and related information • Field edit • Organize image files • May be critical to business communications • Critical to social network and social communications</td>
</tr>
</tbody>
</table>
Table 1. Understanding the devices, practices, skills, and motivations of the digitally enabled traveler
(continued)

| Digital Video Camera | Direct connect to laptop or other devices with Firewire. | Easy to expert level skill | • Record motion and sound
• Record technical file type and related information
• Field edit – shoot to edit
• May be critical to business communications
• Critical to social network and social communications

| Computing Tablet | Direct connect to laptop or other devices with Firewire. Essentially same connectivity as laptop, yet these devices are relatively new to the marketplace and not widely adopted by travelers | Moderate to advance, able to detect wireless signal; may require some basic problem-solving. Application base still developing | • Typically less functional than laptop
• Consumers are urged to match intended use to tablet functionality

Digitally enabled traveler, contact with the social network on the Web is critical. This technical and artistic practice is easy to achieve and among the members of the digital generation, a routine and necessary activity.

Table 1 offers a way to understand the devices, practices, skills, and motivations of a digitally enabled traveler.

Professionals in travel and tourism should understand that the focus of these technologies is the individual consumer, not necessarily the business enterprise.

**Digital Imaging Trek: A Model for the Tourism Industry**

The digitally enabled traveler is equipped with an endless array of digital products designed to capture and record the tourism experience. Through these devices travelers are highly motivated to stay connected to virtual communities online. Patrons of tourism will use their technology skills to create visual and rich media collateral—high quality digital media artifacts. There are countless Web-based outlets to connect the traveler’s experience and collateral to an eager virtual community and social interaction network. As digitally enabled tourism becomes an expectation among patrons and customers, tourism professionals need to understand and shape usage models.

The model of the digital imaging trek proposes information and technical architectures to capture travel experience, create a virtual record of tourism, and meet the demands of the digitally enabled traveler.

From early times of travel, the notion of a “trek” has long been regarded as a journey of self-exploration for the traveler. Many tourists
and travelers today are seeking a heightened experience as part of their tourism through digital communication technologies. Putting practice into models, especially the practice of the digital imaging trek, is a way to understand the processes and practices of the digitally enabled traveler (See Figure 3). Using all the digital tools available, a digital trekker will produce files from a wide range of sources and self-publish material in two methods; saved as CD-based files (or some other permanent memory such as DVD or portable USB Flash Memory) or published on the World Wide Web.

A typical day on digital imaging trek starts with a destination selection. Digitally enabled travelers will turn to the Internet to locate and decide upon a location. Whether it is a museum, a regional historic location or a cultural performance, the digitally enabled traveler will seek all types of information from Web site based services to make plans. When made available on a Web based system, transportation schedules, phone numbers, hours of operations, special information about exhibits, costs, and other related information about the region, are always instantly available to the traveler. Web based systems allow the digitally enabled traveler to acquire information regardless of time and in a preferred language.

In digital imaging trek model, the traveler is focused on acquisition of digital still and video

Figure 3. The general digital imaging trek model embraces technology communication devices and self publishing
Digital Imaging Trek

images that record the daily experience of the trip. In any environment—built or natural—the visitor will always encounter new scenes that are novel, exciting, and educational and from the perspective of the traveler, representative of the tourism experience. With cheap recording memory in the form of Flash Memory, a digitally enabled traveler has near infinite storage space to capture images. Digitally enabled travelers also can edit on the fly. The advantage of digital imaging over traditional film imaging is in the use of storage and field editing. If an image is not worth keeping, it can be immediately erased. Images worth keeping are filed and stored until needed later.

As the day progresses on the digital imaging trek, travelers will want to spend time reviewing images, editing images, and corresponding to the social network with e-mail, uploads and blogs. Editing and organizing images are tasks that most travelers will conduct when time is available. Typically image editing is done on a portable laptop computer. Many people prefer to travel with laptop computers as they are light weight, highly functional, and a repository of software tools and information necessary to conduct digital imaging while traveling. (For the business traveler, a laptop is practically required gear for the trip.) Editing image files requires some time and concentration and will like occur during a break in activities. When editing still images with a popular image editing software such as Adobe Photoshop, travelers will crop, color-correct, merge and manipulate images.

Figure 4. Students and faculty from Johnson & Wales University on a digital imaging trek to Paris; wireless digital connection to the Internet is abundant in urban environments; checking e-mail and editing images in a street side café is now a common tourist experience
Similarly, digital video footage also will be edited. Travelers on a digital imaging trek will likely download and edit video footage with commonly available video editing software such as Adobe Premiere or Apple’s iMovie and Final Cut Pro. Still image editing software and video editing software is widely available. With basic skill and proficiency, digital travelers can achieve remarkably high quality results. For travelers on a digital imaging trek, it is the primary focus of experience and a rewarding achievement to acquire these images.

The ease of use of digital imaging equipment—still cameras, video cameras and laptops—promotes the phenomena of “hyper-imaging”—taking thousands of images to sort through later. Travelers who are serious about digital imaging are continuously shooting images throughout the day and night. The result of hyper-imaging is an overabundance of images and footage that must be sorted and organized. Powerful laptop computers with optimized internal and external storage are perfect for this task. Sorting and organizing images and footage is typically done to suit the desires of the traveler on a digital imaging trek. Images are categorized by group or class and notated with keywords for access at a later time. Various gallery and filing software makes the task of organizing relatively easy.

A Short Primer on Digital Photography

Unlike film cameras, which store images on film, digital cameras capture images via electric sensors and store those images on reusable solid-state memory.

The first digital cameras targeted toward consumers were released in the early 1990s. Since then the digital camera revolution has taken off, and there are hundreds of models available. Competition has remained fierce and manufacturers are producing digital cameras with limitless capacity. Just 15 years after the introduction of the first consumer digital cameras, current models are take pictures that are as high or higher quality than film cameras. Digital cameras take some getting used to, but tourists and travelers who taking digital imaging seriously will find numerous camera choices in the marketplace.

Digital cameras fall into three general categories: consumer, prosumer, and professional. Consumer cameras generally have a single, non-removable lens that mimics a standard 35mm film lens. However, manufacturers now produce consumer level models with nearly every feature of a traditional Single-Lens Reflex or SLR camera.

The term prosumer is a blend of “professional” and “consumer” and refers to consumers who demand more than the standard technology available while being unable to afford professional equipment. In the digital camera market, the prosumer devices are Single-Lens Reflex, or SLR, models, and are falling below the $1000

Figure 5. Digital cameras are popular among travelers and tourists; they are portable, easy to use, and affordable.
price point. Prosumers tend to be technologically savvy and more tolerant of quirks and bugs in new products.

Many professional photographers are moving entirely to digital photography, enticed by the low cost of shots, ease of printing, high resolutions, and ability to edit photos easily with software such as Adobe Photoshop. Professional cameras generally accept the same standard lenses as film cameras and store photos in RAW format, which allows for greater flexibility in editing. The RAW format can be thought of as a digital negative.

The editing and sorting activity is often a precursor activity to the act of e-mail, blogging and posting images on a Web site to interact with digitally connected social network. Travelers with a laptop and access to the Internet will use the online connection to communicate frequently and for sustained periods of time—sometimes hours at a time. If connectivity is available from a wireless service, the traveler with a laptop will connect, log in to services, and communicate on a frequent basis throughout the day. If the Internet connection is more concentrated, in an Internet café with computers for instance, the traveler will dedicate a portion of the day to connect and communicate online. It is not unusual to see Internet cafés through Europe and Asia, but the model is less attractive in the U.S. Internet connection services should allow the traveler to connect and spend as much time as necessary to conduct the typical activities as blogging, image

**Figure 6. Digital imaging trek file types, platforms, and purpose**

<table>
<thead>
<tr>
<th>File Type</th>
<th>Hardware Platform</th>
<th>Software Platform</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text file</td>
<td>Laptop, desktop, or handheld device with keyboard. Privately owned, or available through internet cafe</td>
<td>Word processor-editor software</td>
<td>Written log of experience. Creating and sending attachments on email. Creating and posting blog on web services</td>
</tr>
<tr>
<td>Still image</td>
<td>Digital still camera Cell phone with still camera function Multifunction still and video camera</td>
<td>Embedded software in camera platform. Image editing with additional software on laptop or other device</td>
<td>Create a visual record of still images of travel experience. Sort and publish on web server to share with social network. Send as attachments in email.</td>
</tr>
<tr>
<td>Motion images</td>
<td>Digital video camera Cell phone with motion camera function Multifunction still and video camera</td>
<td>Embedded software in camera platform. Editing with additional software on laptop or other device</td>
<td>Create a visual record of motion images and sound. Edit and publishing on web site to share with social network. Send attachments.</td>
</tr>
<tr>
<td>Hyperlink</td>
<td>Web based hosting services – social networks</td>
<td>Embedded feature on web site</td>
<td>Allows author to ‘link’ from one type of information to another. Series of links create trails for others to follow.</td>
</tr>
</tbody>
</table>
posting, e-mail, checking itinerary, and so on. As the day winds down on the digital imaging trek, the digitally enabled traveler has acquired a new database of images, published and shared select image-files, and communicated stories of their experience with their social network online through e-mail and blogs.

Taking pictures on a vacation is not new to the world of tourism, but digital equipment has changed the business and behavior equation. Given the proper technical infrastructure, images can be acquired and posted—essentially published—to a world-wide audience in a matter of minutes.

Digital photos are easy to share. Most photos are already stored on a computer or memory card, and it is a simple matter to upload them to the Internet, either through a private gallery (such as the digitaltrek.org site), or a social photography Web site such as Flickr (www.flickr.com).

Most thriving urban centers and developed destination resort complexes have a mature ICT infrastructure to support instant access to the Internet. Even smaller rural locations now boast access to the greater world of the Internet. The goal of a traveler on a digital imaging trek is to capture and publish images that reflect travel experience in new places and foreign cultures.

Tourism and Information Architecture

It is critical that both traveler and tourism enterprise alike prepare for the use of digital technologies through “architecture.” Travelers and organizations need to plan and think ahead of how to respond and conduct activities through the travel and tourism experience. While some things are left to chance or serendipity the deployment of digital infrastructure requires some thought. That plan is called “architecture.”

Architecture for information services is similar to that of architecture for buildings—it is the deliberate planning, modeling, and delivery of ICT services. Architecture requires adopting a model of approach and solution, both in systems architecture and information architecture. How does a highly connected, fast speed digital environment that serves the digitally enabled traveler sprout up? It is not a singular act by any tourism organization, but a system wide policy approach of local and national partners in business and governmental bodies. In most cases, experts from companies that provide ICT services will consult on the array of services available in the technology marketplace.

Much of the communications infrastructure that has developed in the past decade is a confluence of commercial profit-driven interest, digital user culture, and regional policy and laws. The combined effect of the built ICT environment serves residents and visitors alike. Localities with new digital infrastructure enjoy digital connection to a broader world. Travelers to areas that are served with high speed digital services also reach out and use the infrastructure for all their related travel communication activities. This ever growing presence of widely accessible digital communication services represents different challenges to different cultures. While most Western cultural sensibilities accept the openness of access to the World Wide Web, other cultures are more circumspect and reserved, wishing to control the vast onslaught of information pouring out of the Web.

Cost has always been considered the primary barrier for entry in the world of ICT. The receding cost barrier—as Moore’s Law predicts—has lowered the cost barrier, particularly with technology products and services. Tourist properties such as hotels and resorts now can achieve Internet access as easily as telephone access. In many instances, travel infrastructure such as airports, train stations, and the aircraft and trains themselves, provide Internet and advanced telephone access. Countries and regions are now adopting
ICT’s and installing a fiber-optic telecommunications system with widespread wireless digital access points, which has distinct advantages over older, legacy “wire based” telecommunications infrastructures—particularly for the traveler. In some ways, the new fiber and wireless based connectivity infrastructure offers advantages for the digitally enabled traveler.

Of course, this entire ICT infrastructure model represents enormous business opportunity for the properties in which they exist. Hotels, resorts, travel destinations, travel hubs, restaurants, coffee houses, business outlets can differentiate themselves from competitors by providing ICT for a fee to the traveler. Travelers in turn, view access to ICT as a cost associated with travel and plan to spend money to acquire it. It is common for travelers to purchase short term access to ICT services at these sites on terms ranging from a few hours to several days.

When collaboration on planning and development of digital infrastructure occurs, the stage is set to promote services to the digitally enabled traveler. A highly developed infrastructure and usage model promotes rapid creation and publishing of rich media collateral by the digitally enabled traveler. All parts of the technology infrastructure work in unison to support the activities of the traveler.

Figure 7. The architecture for wireless network and infrastructure model for digitally enabled traveler
Digital communication technologies of many platforms have become accessible to even the most inexperienced traveler. Mobile technologies of all types and sizes and wireless network signal have brought the Internet to the most distant of locations. This convergence of technical infrastructure and tourism presents opportunities and challenges to the tourism industry.

Yet even as the cyber-record of digital travel experience explodes, there are many issues the profession must consider. The more available common digital technologies become, the more travelers and tourism enterprises grapple with the complexity (and confusion) of choices including but not limited to privacy, intellectual property, systems integration, systems management, training, best practice, and so on.

**Digitally Enabled Travel: Knowledge Targets for the Novice**

As the digitally enabled traveler becomes more of an influence in the world of travel and tourism, young professionals seeking degree based education and career opportunities in the field need to be familiar with basic technology. Whether looking at a college level curriculum or training opportunities for adults wishing to gain new skills, look for some of the topics covered here.

College level curriculum in hospitality, travel, and tourism programs should include an introductory level course in this area of digital technology. The topics taught in such a course do not necessarily have to be technical but young professionals entering the business should know the basics and be prepared to research and understand technology. A sample of college level curriculum might include topics such as the following.

**Foundations in Technology**

Many professionals and educators mistakenly think that learners know all there is to know about basic computer operations, productivity software and Internet access, Web searching and surfing. While these topics are often part of many school systems, not everyone has mastered basic skill and techniques in all these areas. Interacting with computer interfaces and mastering the sophistication of some productivity tools such as word processing, spreadsheets, databases and browsers requires instruction, time and practice. Many colleges, universities, technical institutions and similar educational organizations offer courses in computer technology. The popular press is filled with instructional books complete with CD based video for those who are adept at self instruction. Many conferences, training seminars, and “Webinars” (seminars of text, sound, and motion hosted on the Web), offer similar instruction. Regardless of the model one uses to get trained, using digital technologies requires constant “tune ups” to one’s skill set. Find the type of education that works for you and take advantage of it.

**Basic Information and Systems Architecture**

Information and systems architecture introduces models to help solve problems in the field of technology. As with providing digital services for travels to use to communicate with social networks, technology applications start as a way to provide a solution to a problem.

While it is convenient to say technology will solve a problem or make some situation more efficient or less cumbersome, the idea has to be communicated in a visual model. Professionals in the world of information technology communicate regularly with visual models that reflect the structure or architecture of a system. Devices and the networks that connect them are shown so all can understand and agree on system solutions. Understanding basic architecture in work flow, information flow, and management, networking, service devices, client side devices, and so on will be very valuable for new professionals in the world of hospitality, travel, and tourism.
Digital Imaging Trek

Basic Web Design and Interface

Many professionals in the field of hospitality, travel, and tourism will likely be asked to participate in design teams to design and build Web sites that will host information for customers or be used internally as sources of information for the business enterprise. Conceptualizing and building a Web site is a skill that requires practice and experience although of software tools such as Microsoft FrontPage and Adobe (formerly Macromedia) Dreamweaver make it easy to try this activity. With a little training in software features of such applications, college level courses can quickly teach students the essentials in basic Web site design. This is both a technical and creative challenge, but most people quickly see the results of a little effort. For people just wishing to start with a simple Web site to host information about a trip or a destination, these tools are cheap, accessible, and relatively easy to learn.

Basic Digital Imaging and Image Manipulation

There are an abundance of off-the-shelf software tools in the marketplace that will introduce college student (or anyone with the motivation to learn) skills in digital imaging and image manipulation. While many people have digital cameras a course on digital photography will introduce a wide range of topics from composition to technical specifications. Understanding the wide range of options in today’s digital cameras can be helpful for young professionals who need to deal with a customer base that is armed with the latest camera gadgetry. Learning the process involved in capturing an image and uploading an image to a devices that can store the image in important. More knowledge in these areas will help improve customer empathy and ultimately customer satisfaction.

Editing images in software like Adobe Photoshop has a double benefit to a college curriculum. Students with added technical skills such as digital photo editing are in more demand. Many businesses need such skills to help with developing creative collateral to promote the commercial activity of the business. Understanding the creative process from image acquisition to image editing also will help the professional in dealing with agencies that provide that service. For instance, image editing is a time consuming activity, and in negotiating advertising contracts, such knowledge will be helpful.

Desktop Publishing

Desktop publishing is a content area that will provide a young professional with the knowledge of how to put assets such as text and images together in print collateral to serve the business. Even in the digital age, many businesses still have a great need for printed products to advertise, inform and attract potential customers. Tourism properties still have a great need for pamphlets, brochures and signage to keep customers informed about policies, regulations, events, calendars, and so on. A course in desktop publishing will give practitioners an opportunity to learn a skill that will help promote and organize business. In desktop publishing, students can learn how to conceive and construct various types of print pieces using software such as Microsoft Publishing or Adobe Indesign.

Editorial, Content Creation and Content Management

While learning technical tools is vital, helping young professionals identify and manage the message is critical to the success of travel and tourism businesses. Courses that emphasize the basics of how to construct the message in both text and visual design are important for a basic college curriculum.
Systems and Technology Primer for the Digitally Enabled Traveler

As digital communications platforms and technologies have become adopted by a wide mainstream audience two factors have been critical to widespread use—practicality and ubiquity. Along with the explosion of digital gadgets, contemporary travelers now have high expectations of availability of connectivity and complimentary technologies.

If the tourism industry is to appropriately harness and cater to this new and demanding audience, the industry must build a model of digitally enabled travel that supports both traditional goals of tourism and the new goals of digital media acquisition. Understanding the pieces of the technology puzzle can serve as a starting point.

Several recent world events have highlighted the role mobile technology plays on the world stage. The first images of the London tube bombing in 2005 came from the cell phone cameras of survivors. These images were published by the BBC and forwarded around the world within minutes of the attack. Similarly dramatic images, particularly digital video, were quickly spread after the 2005 tsunami disaster in the Indian Ocean. Whether it is a global or local scale, digital imaging devices are ever present and serve as eyes to the world. Building a world class digital environment brings with it many more challenges, not just in the technical realm, but in the human realm as well.

A Word on Privacy and Security in the Digital Age

Privacy, security of information, copyright, information ownership, censored, and uncensored material are all issues that become concerns of the industry when technology is introduced. It is a grey area because decorum and respect relies as much on personal behavior as it does on personal technology. Travelers with powerful digital recording tools must understand the local cultural norms, as well as the global broadcast power of the World Wide Web.

Privacy is an important and sensitive issue. Visitors to a Web site may be reluctant to share information—such as their e-mail address—because they fear that their personal information will be sold and they will be subjected to unsolicited advertisements. Spam, the endless barrage of meaningless e-mail advertising, has become an onerous burden to all citizens using online tools. It is best avoided by constricting the use of e-mail addresses.

Information that may be personal or sensitive, such as vacation photos or a travel journal, should be posted with care. In many cultures people do not wish to have their images posted for the world to see. Privacy for individuals is a sensitive matter. Even in public tourism venues, digital photographers are challenged when taking images and asked to refrain. Religious and private properties often post requests to refrain from photography and video recordings.

Institutional Web sites should consider developing a privacy policy, or a legal statement that reflects what the institution may and may not do with information provided by users. The World Wide Web Consortium (the standards body for the Web) has developed the Platform for Privacy Preferences (P3P) (http://www.w3.org/P3P/), a standardized language which provides Web site administrators with a simple and automated way to quickly generate a customized policy for their site. Individual digital travelers, eager to capture a unique image, must apply their own standards. The golden rule though is “do unto others” as you would have done to yourself. Consider the impact of each image that is posted.

Travelers must be aware of their physical safety and security at all times. Broadcasting information on the Web can add to this worry. For example, travelers, especially solo-travelers, are advised
against posting a personal and complete itinerary until their trip is complete. No only does it tell the world where you will be, but it also tells the world where you are not. Home safety as well as trip safety is the paramount concern.

With reasonable precautions, a Web site can increase safety. It allows a large number of people to check in on the well-being of travelers. While theft and other dangers cannot be eliminated, careful use of a Web site, instant messaging, and e-mail can reassure those back home. The personal technology of a digital enabled traveler is an attractive target for thieves. Common sense should guide the novice and experienced traveler in protecting personal possessions.

### Web Hosting: What Travel and Tourism Professionals Need to Know

All Internet applications such as a Web site or a search engine run on a server. The server is a computer workstation in a class of computing machines that is specially constructed to manage the constant demand of service to clients on the network. Servers require special software and in most cases, comparably advanced knowledge and skill to configure and maintain. While it is possible to run a server in-house, it involves purchasing a machine, installing and supporting an operating system and applications, and maintaining an “always-on” Internet connection. Managing a server

<table>
<thead>
<tr>
<th>Technology</th>
<th>Location</th>
<th>Skill-level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Client</td>
<td>Computing cluster such as an Internet café, business service office, hotel lobby, etc. One or more for use by customers</td>
<td>Basic to advanced software such as email, word processing, file management; basic computer knowledge</td>
<td>The computer client serves as the starting point for the customer. Recreational and business travelers alike will budget time and funds to access basic computing services such as email, word processing, file uploading.</td>
</tr>
<tr>
<td>Laptop Computers</td>
<td>Anywhere, travelers port and manage</td>
<td>Basic to advanced software such as email, word processing, file management; basic computer knowledge</td>
<td>A laptop computer is a computer client and the traveler’s link to the internet and workstation for writing blogs or editing digital photos.</td>
</tr>
<tr>
<td>Card Readers</td>
<td>Anywhere.</td>
<td>Basic computer knowledge required.</td>
<td>Digital cameras accept different memory cards depending on brand. A generic card reader will allow any computer to read any card from any camera, with no additional software.</td>
</tr>
</tbody>
</table>

Table 2. Core components of an information technology system
Table 2. Core components of an information technology system

<table>
<thead>
<tr>
<th>Technology</th>
<th>Location</th>
<th>Skill-level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portable Storage</td>
<td>Anywhere</td>
<td>Basic computer knowledge required.</td>
<td>Backups are a major issue for travelers concerned about potential data loss (from theft or equipment failure). A portable hard drive (or a device that can burn CD’s) provides a cheap and easy backup solution on the road. Small and easily packed.</td>
</tr>
<tr>
<td>Network – wired</td>
<td>Anywhere</td>
<td>Intermediate to advanced computer knowledge required to build and support. Entry level user knowledge to access and use.</td>
<td>Wired internet connections are generally located in institutions (such as hotels, business services, universities), and may incur a cost for access.</td>
</tr>
<tr>
<td>Network -wireless</td>
<td>Anywhere</td>
<td>Intermediate to advanced computer knowledge required to build and support. Entry level user knowledge to access and use.</td>
<td>Wireless internet connections can be found in built environments, particularly popular in urban settings with restaurants, parks, museums, hotels, and cafes.</td>
</tr>
<tr>
<td>Mass Storage</td>
<td>Any secure building location, typically attached to host devices such as advanced workstations or mainframes.</td>
<td>Intermediate computer knowledge required to configure, install and support.</td>
<td>A mass storage device, is a very large, commercial grade hard drive. It supports data and functions core to large businesses. While computer users may see the results of such devices in a routine web search, these devices work in the background of daily user activities.</td>
</tr>
<tr>
<td>Servers</td>
<td>At a hosting company; typically a technology company which provides a secure physical location, as well as technical knowledge to support services. Locations are typically built specifically to house the servers.</td>
<td>Advanced technical skills in web hosting and server based data services.</td>
<td>The server will host all of the content for a web site and code for web applications such as a gallery or blog. Often a class of computers known as workstations; function solely as servers. A hosting service provider will maintain and support all server hardware and software. Computer users access these devices through client computers to update websites, post images in galleries and blog.</td>
</tr>
</tbody>
</table>
Digital Imaging Trek

requires routine management as well as prompt response emergencies, 24 hours a day, 7 days a week. For most travel and tourism providers, becoming a technology company is a distraction from the core business. It is generally advisable that any business requiring a host on a server be outsourced to a Web host business provider.

A Web host is a company that provides space on their Web servers, use of their programs, and a certain amount of bandwidth use for a monthly cost. Often referred to as a “solutions provider” or a “Web hosting service” the hosting company will handle all computer hardware and software issues such as installations, configurations, back-ups, updates, and any unforeseen maintenance. Of course, solution providers contract these services for a fee, but competition works in this marketplace as it does in any other, and a shopper is wise to research and compare costs and services among a range of businesses.

There are many Web hosting companies and there are many attributes to consider when researching them. Choosing a hosting company is a long-term commitment. Transferring a site from one hosting service to another generally involves several days of unreliability and possible downtime. Carefully consider all of the factors outlined in Table 3 before making a decision.

All of these considerations are important indicators of a successful Web hosting experience. Remember, anyone can run a server, including a college student living in an apartment with a space for a computer on a network. Make sure that you are dealing with a professional company that has

Table 3. Qualities for evaluating a hosting company

<table>
<thead>
<tr>
<th>Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reputation</td>
<td>Every company must market itself on its reputation for service to customers. Ask for reference of other businesses that have used the services. Search for comments on the web about the company and its performance. Are there any instances of how the company performed in a crisis, such as a virus attack or power failure? Have you toured the facilities or conducted due diligence on the company?</td>
</tr>
<tr>
<td>Reliability</td>
<td>Does the company publish statistics on technical performance? What level of support and customer service contact is there? Will you be able to contact people in off hours? What kind of technology do they support? Is there a specific platform or hardware and software and do you recognize the vendor? Ask about back-up, power supply, physical security, data security, etc. Do they publish their reliability data? Conduct an internet search through popular search engines to find information about them.</td>
</tr>
<tr>
<td>Technology</td>
<td>Is the hardware and software platform up to date and current? Can you speak to vendors who supply the company with technology. Do they publish any related information about platforms?</td>
</tr>
<tr>
<td>People</td>
<td>Making a business agreement is as much about people as it is price structure. Have you met the principals of the company? Are you generally familiar with their organization and business structure? Do you have confidence in the people you have met? Is there generally superior business communication to your proposals?</td>
</tr>
</tbody>
</table>
a reputation for handling routine business as well as crises. Like the travel and tourism industry, professional protocol in handling technical matters as well as high grades on customer service are key to sound business practice in the technology marketplace.

As with every business, hosting is about people as much as technology. Find a company with knowledgeable employees who are happy to deal with their clients. You should find support personnel who are professional, courteous and eager to answer questions at your knowledge level, no matter how basic or advanced.

Every good host publishes their uptime (the amount of time the service has run without interruptions) and customer testimonials. Don’t necessarily trust the quotes on the company’s Web site; do an Internet search with any popular search engine (Yahoo, Google, etc.), and find out for yourself.

Once you have identified a few companies that you feel comfortable dealing with, consider the technical requirements of your project. The first decision is about the type of hosting that the project requires. The options are:

- **Dedicated hosting** means that you are leasing an entire computer, which gives you access to the entire hard drive and allows you to make certain configuration requests that are not possible on a shared machine. By leasing a server (rather than purchasing it and placing it in your office), you outsource the need for physical setup, administration and backup services. You also avoid the upfront cost of a powerful machine, and take advantage of your hosting company’s ability to quickly purchase and maintain the machine with professional qualified vendors.

- **Shared hosting**, sometimes called managed hosting, means that you are leasing part of a server, which you will share with other customers. This kind of plan generally involves very low monthly fees (as low as $10/month) and provides complete technical support. Unless you are hosting multiple blogs or Web sites, a shared hosting plan is probably sufficient.

Travel and tourism organizations could have a range of needs to contract an outside hosting provider. Setting up a Web page to inform customers of features and offerings of a property or package, or architecting more advanced online services for customers may be among the ideas you will have. The next step in this process of selecting a hosting provider is to explore what you want to do with your hosting provider. The type of service you offer will determine the software applications that the hosting provider will need to run on the server. Additionally, your requirements will guide the features that you’ll select for the server. Each software package will list the special requirements on their Web site, but some general guidelines are:

- **Static Web page**: Simple Web pages that hyperlink information that is relatively static. Information does not change too often. Static Web pages require little to no special software or technical support.

- **Downloadable images**: No special requirements, similar to static Web page

- **Image gallery with uploading**: Image galleries containing collections of image (often hundreds and thousands) require server-side programs. Server side programs mean that special software has been loaded and manipulated to provide an easy user interface and experience. A database will be required to track and maintain a medium to large size collection of gallery images.

- **Blogs**: Blogs, or written text entries also require special server-side software to accepted entries, organize them and display them, blog software, like image galleries is generally database-driven.
Several hosts offer blog-friendly plans that require minimal configuration and no technical skills. Lists of these can be found at http://wordpress.org/hosting/ and http://www.sixapart.com/movabletype/hosting.

Identify your minimum technical requirements. In general, look for the criteria listed in Table 4.

**More Technical Tools of the Digitally Enabled Traveler**

Digitally enabled travelers will quickly embrace new and experimental technologies. For them, it is a challenge to figure out ways to adopt a technology and make it work in the daily flow of travel.

For instance, instant messaging, or “IM,” is a technology for rapid-fire asynchronous messaging across a network. Instant messages can be thought of “instant e-mail,” and are generally used to communicate to a select social network, but often composed on the computer screen while doing something else. Digitally enabled travelers multitask on the computer—that is, they conduct more than one activity at the same time—with ease.

<table>
<thead>
<tr>
<th>Feature</th>
<th>What to Look For</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical Support</td>
<td>At the very least, 24-hour email technical support. Also consider telephone support (generally, the company promises to address your problem and call you back within 24 hours).</td>
</tr>
<tr>
<td>Domain</td>
<td>Most hosting plans include one domain registration (for example, <a href="http://www.yournamehere.com">www.yournamehere.com</a>) for free. Make sure that the price for a second is reasonable (not more than $20/year).</td>
</tr>
<tr>
<td>Disk space allocation</td>
<td>Disk space, the space on the hard drive that is allocated to you, should be adequate for your application. Storing web pages or blogs requires a small amount of space, while storing photographs requires multiple gigabytes. Look for at least 10GB of space. Also, research in advance what it will cost to expand your allocated space. Understand the incremental hikes in cost for more disk space.</td>
</tr>
<tr>
<td>Server-side program access</td>
<td>The applications that you wish to run will guide your need for server-side program access. In general, look for PHP, which is a coding language that manages how customers see and interact with your services.</td>
</tr>
<tr>
<td>E-mail addresses</td>
<td>Your plan should offer the ability to create your own <a href="mailto:me@mydomain.com">me@mydomain.com</a> e-mail addresses. Web-based e-mail access is a nice feature for frequent travelers, and is offered at no additional cost by many hosts.</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>Bandwidth refers to the total amount of data transfer between your server and visitors to your web site. More bandwidth is good. Understand that bandwidth to the end user is limited by their personal connection and rate of speed.</td>
</tr>
<tr>
<td>Statistics</td>
<td>It’s always helpful to know exactly how many visitors your site has. Hosting companies typically provide this information with some graphics and analysis.</td>
</tr>
<tr>
<td>Databases</td>
<td>If you wish to run a dynamic application such as a blog, you’ll need access to a database.</td>
</tr>
<tr>
<td>Operating System</td>
<td>Your operating system choice should be guided by your application requirements. Most hosting plans are based on either Linux (which tends to be cheaper), or Windows.</td>
</tr>
</tbody>
</table>
IM was first popularized in the mid-1990s by ICQ, a product introduced by Mirabilis Ltd in 1996. Mirabilis Ltd was later acquired by America Online, which ran a competing product called “AOL Instant Messenger,” or AIM. By acquiring ICQ, AOL became the largest operator of instant messaging networks in the world.

Other competitors include Microsoft, which offers an instant messaging feature as part of the Microsoft Network (MSN), and Yahoo!, which offers the feature as part of their community-building options. All of these services are free to use, though some may display advertisements.

Unfortunately, all of these services are based on proprietary protocols and do not interoperate. Therefore, a digital traveler must choose the service that most of the people that they wish to communicate with use. If this is not practical, digital travelers can invest in a product such as Trillian, from Cerulean Studios (http://www.ceruleanstudios.com/), which is an IM program that supports all major networks.

Instant messaging is favored by digital travelers, because it allows instant communication with anyone on the network, no matter where their physically located. If a device (such as a computer or cellular telephone) has Internet access, it can connect to an IM network. This is an extremely low-cost method for travelers to communicate with their social network of friends, family, and colleagues around the world.

More recently, voice over Internet protocol (VoIP) has become an important buzzword in the latest digital tools becoming available to digital travelers. VoIP is a technology for transmitting traditional “telephone calls” over the Internet. Imagine using a laptop as a phone and conducting routine two-way audible conversations. Such calls are indistinguishable from normal Internet data and promise to make international conversational calls extremely cheap. Imagine no more phone calling cards and the end to frustrated travelers trying to decipher local and international calling codes.

While still in the early stages, VoIP is available to any digital traveler with a laptop, speakers, and a microphone. Services such as Google Talk or Skype (http://www.skype.com/) are up and running and can be used to converse with telephone quality audio for free. Skype also offers “Skype Out,” which permits computers to connect with telephone numbers anywhere in the world for extremely reasonable rates.

Needless to say, it is important for travel and tourism organizations to be familiar with trends such as VoIP for two reasons. The early adopters of such technology—the digitally enabled traveler—are the bell weather for change in the industry. It won’t take long for entrepreneurial technology providers to create a competitive business model for VoIP and shift the commerce to a new marketplace.

CONVERGENCE AND FUTURE TRENDS

Given the rapid evolution of digital communication tools, both hardware and software based, it is not hard to imagine that many of these products will continue to transform.

The technology industry is creating faster, lighter, devices that better integrate with each other. This trend of “convergence” is one to watch for. It wasn’t long ago that a cell phone was simply a way to have voice conversations with other parties. Today, other features have converged on the platform of a cell phone. Is a cell phone a camera, a video recorder, a video viewer, organizer and personal digital library? The answer is “yes”—it is all those things and more.

A digital camera is no longer a camera; it is an image processing workstation. Images can be acquired, edited, filed, and stored. Images are dynamic files that merge and morph into other applications and devices through infrared proximity connections.
The ubiquitous iPod from Apple computer started off not too long ago as a just another platform for listening to music. Today an entire lucrative industry revolves around providing rich media content for viewing on a tiny screen. Need an exercise multimedia instruction package for keeping fit during travel, it can be purchased and downloaded to your iPod.

CONCLUSION

Digital gear to enhance the travel and tourism industry is a constantly evolving marketplace. The future is part evolution and part revolution, from wildly popular devices to culturally challenging information flow. Regardless of the intricacy, allure or popularity of digital devices, the seasoned digitally enabled traveler will always find ways of incorporating new devices and new services to enhance experience and communicate to a social network in digital space. Travel and tourism providers must influence the role of the digital imaging trek by understanding the role they themselves play in this ever changing landscape. Who is the provider, the host, the arbitrator of this new world of digital travel? By seeing and understanding the big digital picture, tourism and travel providers will retain perspective and offer quality services to patrons worldwide.

NOTE

- All images are provided with permission to reproduce by the authors.
- Illustrations are public domain license free illustrations from Microsoft Visio.
- Figures representing models are provide with permission to reproduce by the authors.
- Illustrations in models are public domain license free illustrations from Microsoft Visio.

This work was previously published in Information and Communication Technologies in Support of the Tourism Industry, edited by W. Pease, M. Rowe, and M. Cooper, pp. 286-313, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Chapter 8.6
Intelligent User Interfaces for Mobile Computing

Michael J. O’Grady
University College Dublin, Ireland

Gregory M. P. O’Hare
University College Dublin, Ireland

ABSTRACT

In this chapter, the practical issue of realizing a necessary intelligence quotient for conceiving intelligent user interfaces (IUIs) on mobile devices is considered. Mobile computing scenarios differ radically from the normal fixed workstation environment that most people are familiar with. It is in this dynamicity and complexity that the key motivations for realizing IUIs on mobile devices may be found. Thus, the chapter initially motivates the need for the deployment of IUIs in mobile contexts by reflecting on the archetypical elements that comprise the average mobile user’s situation or context. A number of broad issues pertaining to the deployment of AI techniques on mobile devices are considered before a practical realisation of this objective through the intelligent agent paradigm is presented. It is the authors hope that a mature understanding of the mobile computing usage scenario, augmented with key insights into the practical deployment of AI in mobile scenarios, will aid software engineers and HCI professionals alike in the successful utilisation of intelligent techniques for a new generation of mobile services.

INTRODUCTION

Mobile computing is one of the dominant computing usage paradigms at present and encapsulates a number of contrasting visions of how best the paradigm should be realized. Ubiquitous computing (Weiser, 1991) envisages a world populated with artefacts augmented with embedded computational technologies, all linked by transparent high-speed networks, and accessible in a seamless anytime, anywhere basis. Wearable computing (Rhodes, Minar, & Weaver, 1999) advocates a
Intelligent User Interfaces for Mobile Computing

world where people carry the necessary computational artefacts about their actual person. Somewhere in between these two extremes lies the average mobile user, equipped with a PDA or mobile phone, and seeking to access both popular and highly specialized services as they go about their daily routine.

Though the growth of mobile computing usage has been phenomenal, and significant markets exists for providers of innovative services, there still exist a formidable number of obstacles that must be surpassed before software development processes for mobile services becomes as mature as current software development practices. It is often forgotten in the rush to exploit the potential of mobile computing that it is radically different from the classic desktop situation; and that this has serious implications for the design and engineering process. The dynamic nature of the mobile user, together with the variety and complexity of the environments in which they operate, provides unprecedented challenges for software engineers as the principles and methodologies that have been refined over years do not necessarily apply, at least in their totality, in mobile computing scenarios.

How to improve the mobile user’s experience remains an open question. One approach concerns the notion of an application autonomously adapting to the prevailing situation or context in which end-users find themselves. A second approach concerns the incorporation of intelligent techniques into the application. In principle, such techniques could be used for diverse purposes, however, intelligent user interfaces (IUIs) represent one practical example where such techniques could be usefully deployed. Thus the objective of this chapter is to consider how the necessary intelligence can be effectively realized such that software designers can realistically consider the deployment of IUIs in mobile applications and services.

BACKGROUND

Research in IUIs has been ongoing for quite some time, and was originally motivated by problems that were arising in standard software application usage. Examples of these problems include information overflow, real-time cognitive overload, and difficulties in aiding end-users to interact with complex systems (Höök, 2000). These problems were perceived as being a by-product of direct-manipulation style interfaces. Thus, the concept of the application or user interface adapting to circumstances as they arose was conceived and the terms “adaptive” or “intelligent” user interfaces are frequently encountered in the literature. How to effectively realize interfaces endowed with such attributes is a crucial question and a number of proposals have been put forward. For example, the use of machine learning techniques has been proposed (Langley, 1997) as has the deployment of mobile agents (Mitrovic, Royo, & Mena, 2005).

In general, incorporating adaptability and intelligence enables applications to make considerable changes for personalization and customization preferences as defined by the user and the content being adapted (O’Connor & Wade, 2006). Though significant benefits can accrue from such an approach, there is a subtle issue that needs to be considered. If an application is functioning according to explicit user defined preferences it is functioning in a manner that is as the user expects and understands. However, should the system autonomously or intelligently adapt its services based on some pertinent aspect of the observed behavior of the user, or indeed, based on some other cue, responsibility for the system behavior moves, albeit partially, from the user to the system. Thus, the potential for a confused user or unsatisfactory user experience increases.

A natural question that must now be addressed concerns the identification of criteria that an
application might use as a basis for adapting its behavior. Context-aware computing (Schmidt, Beigl & Gellersen, 1999) provides one intuitive answer to this question. The notion of context first arose in the early 1990s as a result of pioneering experiments in mobile computing systems. Though an agreed definition of context has still not materialized, it concerns the idea that an application should factor in various aspects of the prevailing situation when offering a service. What these aspects might be is highly dependent on the application domain in question. However, commonly held aspects of context include knowledge of the end-user, for example through a user model; knowledge of the surrounding environment, for example through a geographic information system (GIS) model; and knowledge of the mobile device, for example through a suitably populated database. Other useful aspects of an end-user’s context include an understanding of the nature of the task or activity currently being engaged in, knowledge of their spatial context, that is, location and orientation, and knowledge of the prevailing social situation. Such models can provide a sound basis for intelligently adapting system behavior. However, capturing the necessary aspects of the end-user’s context and interpreting it is frequently a computationally intensive process, and one that may prove intractable in a mobile computing context. Indeed, articulating the various aspects of context and the interrelationships between them may prove impossible, even during system design (Greenberg, 2001). Thus, a design decision may need to be made as to whether it is worth working with partial or incomplete models of a user’s context. And the benefit of using intelligent techniques to remedy deficiencies in context models needs to be considered in terms of computational resources required, necessary response time and the ultimate benefit to the end-user and service provider.

**SOME REFLECTIONS ON CONTEXT**

Mobile computing spans many application domains and within these, it is characterized by a heterogeneous landscape of application domains, individual users, mobile devices, environments and tasks (Figure 1). Thus, developing applica-

*Figure 1. An individual’s current activity is a notoriously difficult aspect of an individual’s context to ascertain with certainty*
tions and services that incorporate a contextual component is frequently an inherently complex and potentially time-consuming endeavor, and the benefits that accrue from such an approach should be capable of being measured in some tangible way. Mobile computing applications tend to be quite domain specific and are hence targeted at specific end-users with specialized tasks or objectives in mind. This is in contrast to the one-size-fits-all attitude to general purpose software development that one would encounter in the broad consumer PC arena. For the purposes of this discussion, it is useful to reflect further on the following aspects of the average mobile user’s context: end-user profile, devices characteristics, prevailing environment and social situation.

User Profile

Personalization and customization techniques assume the availability of sophisticated user models, and currently form an indispensable component of a number of well-known e-commerce related Web sites. Personalizing services for mobile computing users is an attractive proposition in many domains as it offers a promising mechanism for increasing the possibility that the end-users receive content that is of interest to them. Though this objective is likewise shared with owners of e-commerce sites, there are two issues that are of particular importance when considering the mobile user. Firstly, mobile interactions are almost invariably short and to the point. This obligates service providers to strive to filter, prioritize, and deliver content that is pertinent to the user’s immediate requirements. The second issue concerns the question of costs. Mobile users have to pay for services, which may be charged on a KB basis, thus giving mobile users a strong incentive to curtail their use of the service in question if dissatisfied.

A wide number of features and characteristics can be incorporated into user models. As a basic requirement, some information concerning the user’s personal profile, for example, age, sex, nationality and so on, is required. This basic model may then be augmented with additional sub-models that become increasingly domain-specific. In the case of standard e-commerce services, a record of the previous purchasing history may be maintained and used as a basis for recommending further products. Electronic tourist guides would require the availability of a cultural interest model, which as well as indicating cultural topics of interest to the user, would also provide some metric that facilitated the prioritization of their cultural interests.

Device Characteristics

Announcements of new devices are occurring with increasing frequency. Each generation successively increases the number of features offered, some of which would not be associated with traditional mobile computing devices, embedded cameras and MP3 players being cases in point. Though offering similar features and services, there are subtle differences between different generations, and indeed interim releases within the same generation, that make the life of a service provider and software professional exceedingly difficult and frequently irritating. From an interface perspective, screen size and support for various interaction modalities are two notable ways in which devices differ, and these have particular implications for the end-user experience. This problem is well documented in the literature and a number of proposals have been put forward to address this, the plasticity concept being a notable example (Thevenin & Coutaz, 1999). Other aspects in which mobile devices differ include processor, memory and operating system; all of which place practical limitations on what is computationally feasible on the device.
**Prevaling Environment**

The notion of environment is fundamental to mobile computing and it is the dynamic nature of prevailing environment in which the mobile user operates that most distinguishes mobile computing from the classic desktop usage paradigm. As an illustration, the case of the physical environment is now considered, though this in no way diminishes the importance of the prevailing electronic infrastructure. Scenarios in which mobile computing usage can occur are multiple and diverse. The same goes for physical environments. Such environments may be hostile in the sense that they do not lend themselves to easily accessing electronic infrastructure such as telecommunications networks. Other environments may experience extreme climatic conditions thus causing equipment to fail.

Developing a service that takes account of or adapts to the local physical environment is an attractive one. Two prerequisites are unavoidable, however. A model of the environment particular to the service domain in question must be available, and the location of the end-user must be attainable. In the former case, the service provider must construct this environmental model, possibly an expensive endeavor in terms of time and finance. In the latter case, an additional technological solution must be engaged—either one based on satellites, for example GPS, or one that harnesses the topology of the local wireless telecommunications networks. Each solution has its respective advantages and disadvantages, and a practical understanding of each is essential. However, by fulfilling these prerequisites, the service provider is in a position to offer services that take the end-users’ physical position into account. Indeed, this vision, often termed location-aware computing (Patterson, Muntz & Pancake, 2003), has grasped the imagination of service providers and end-users alike. In essence, it is a practical example of just one single element of an end-user’s context being interpreted and used as a basis for customizing services.

**Social Situation**

Developing a service that adapts to the end-user’s prevailing social context is fraught with difficulty, yet is one that many people would find useful. What exactly defines social context is somewhat open to interpretation but in this case, it is considered to refer to the situation in which end-users find themselves relevant to other people. This is an inherently dynamic construct and capturing the prevailing social situation introduces an additional level of complexity not encountered in the contextual elements described previously.

In limited situations, it is possible to infer the prevailing social situation. Assuming that the end-user maintains an electronic calendar, the detection of certain keywords may hint at the prevailing social situation. Examples of such keywords might include lecture, meeting, theatre and so on. Thus, an application might reasonably deduce that the end-user would not welcome interruptions, and, for example, proceed to route incoming calls to voicemail and not alert the end-user to the availability of new email. Outside of this, one has to envisage the deployment of a suite of technologies to infer social context. For example, it may be that a device, equipped with a voice recognition system, may be trained to recognize the end-user’s voice, and on recognizing it, infer that a social situation is prevailing. Even then, there may be a significant margin of error; and given the power limitations of the average mobile device, running a computationally intensive voice recognition system continuously may rapidly deplete battery resources.
ARTIFICIAL INTELLIGENCE IN MOBILE COMPUTING

Artificial intelligence (AI) has been the subject of much research, and even more speculation, for almost half a century by now. Though failing to radically alter the world in the way that was envisaged, nevertheless, AI techniques have been successfully harnessed in a quite a number of select domains and their incorporation into everyday applications and services continues unobtrusively yet unrelentingly. Not surprising, there is significant interest amongst the academic community in the potential of AI for addressing the myriad of complexity that is encountered in the mobile computing area. From the previous discussion, some sources of this complexity can be easily identified. Resource management, ambiguity resolution, for example, in determining contextual state and resolving user intention in multimodal interfaces, and adaptation, are just some examples. Historically, research in AI has focuses on various issues related to these very topics. Thus, a significant body of research already exists in some of the very areas that can be harnessed to maximum benefit in mobile computing scenarios. A detailed description of these issues may be found elsewhere (Krüger & Malaka, 2004).

One pioneering effort at harnessing the use of intelligent techniques on devices of limited computational capacity is the Ambient intelligence (AmI) (Vasilakos & Pedrycz, 2006) initiative. AmI builds on the broad mobile computing vision as propounded by the ubiquitous computing vision. It is of particular relevance to this discussion as it is essentially concerned with usability and HCI issues. It was conceived in response to the realization that as mobile and embedded artefacts proliferate, demands for user attention would likewise increase, resulting in environments becoming inhabitable, or more likely, people just disabling the technologies in question. In the AmI concept, IUIs are envisaged as playing a key role in mediating between the embedded artefacts and surrounding users. However, AmI does not formally ratify the use of any particular AI technique. Choice of technique is at the discretion of the software designer whose selection will be influenced by a number of factors including the broad nature of the domain in question, the requirements of the user, the capability of the available technology and the implications for system performance and usability.

Having motivated the need for AI technologies in mobile contexts, practical issues pertaining to their deployment can now be examined.

STRATEGIES FOR HARNESING AI TECHNIQUES IN MOBILE APPLICATIONS

It must be reiterated that AI techniques are computationally intensive. Thus, the practical issue of actually incorporating such techniques into mobile applications needs to be considered carefully. In particular, the implications for performance must be determined as this could easily have an adverse effect on usability. There are three broad approaches that can be adopted when incorporating AI into a mobile application and each is now considered.

Network-Based Approach

Practically all mobile devices are equipped with wireless modems allowing access to data services. In such circumstances, designers can adopt a kind of client/server architecture where the interface logic is hosted on the mobile devices and the core application logic deployed on a fixed server node. The advantage of such an approach is that the designer can adopt the most appropriate AI technologies for the application in question. However, the effect of network latency must be considered. If network latency is significant, the usability of the application will be adversely
affected. Likewise, data rates supported by the network in question must be considered. Indeed, this situation is aggravated when it is considered that a number of networks implement a channel sharing system where the effective data rate at a given time is directly proportional to the number of subscribers currently sharing the channel. It is therefore impossible to guarantee an adequate quality of service (QoS) making the prediction of system performance difficult. Often, the worst case scenario must be assumed. This has particular implications where the AI application on the fixed server node needs either a significant amount of raw data or a stream of data to process.

One key disadvantage of placing the AI component on a fixed server node concerns the issue of cost. There is a surcharge for each KB of data transferred across the wireless network, and although additional revenue is always welcome, the very fact that the subscriber is paying will affect their perception of the application in question and make them more demanding in their expectations.

A network-based AI approach is by far the most common and has been used in quite a number of applications. For example, neural networks have been used for profiling mobile users in conversational interfaces (Toney, Feinberg & Richmond, 2004). InCa (Kadous & Sammut, 2004) is a conversational agent that runs on a PDA but uses a fixed network infrastructure for speech recognition.

Distributed Approach

In this approach, the AI component of the service may be split between the mobile device and the fixed network node. The more computationally expensive elements of the service are hosted on the fixed network node while the less expensive elements may be deployed on the device. Performance is a key limitation of this approach as the computational capacity of the devices in question as well as the data-rates supported by the wireless network can all contribute to unsatisfactory performance. From a software engineering perspective, this approach is quite attractive as distributed AI (DAI) is a mature research discipline in its own right; and a practical implementation of DAI is the multi-agent system (MAS) paradigm.

One example of an application that uses a distributed approach is Gulliver’s Genie (O’Grady & O’Hare, 2004). This is a tourist information guide for mobile tourists, realized as a suite of intelligent agents encompassing PDAs, wireless networks and fixed network servers. Agents on the mobile device are responsible for manipulating the user interface while a suite of agents on the fixed server collaborate to identify and recommend multimedia content that is appropriate to the tourist’s context.

Embedded Approach

As devices grow in processing power, the possibility of embedding an AI based application on the actual physical device becomes ever more feasible. The key limitation is performance, which is a direct result of the available hardware. This effectively compromises the type of AI approach that can be usefully adopted. Overtime, it can be assumed that the capability and variety of AI techniques that can be deployed will increase as developments in mobile hardware continue and the demand for ever-more sophisticated applications increases. From an end-user viewpoint, a key advantage of the embedded approach concerns cost as the number of connections required is minimized.

One example of an application that uses the embedded approach is iDorm (Hagras et al., 2004), a prototype AmI environment. This environment actually demonstrates a variety of embedded agents including fixed motes, mobile robots and PDAs. These agents collaborate to learn and predict user behavior using fuzzy logic principles and, based on these models, the environment is adapted to the inhabitant’s needs.
Deployment Considerations

Technically, all three approaches are viable, but the circumstances in which they may be adopted vary. For specialized applications, the networked AI approach is preferable as it offers greater flexibility and maximum performance, albeit at a cost. For general applications, the embedded approach is preferable, primarily due to cost limitations, but the techniques that can be adopted are limited. The distributed approach is essentially a compromise, incorporating the respective advantages and disadvantages of both the networked and embedded approach to various degrees. Ultimately, the nature of the application domain and the target user base will be the major determinants in what approach is adopted. However, in the longer term, it is the embedded approach that has the most potential as it eliminates the negative cumulative effect of network vagrancies, as well as hidden costs. Thus, for the remainder of this chapter, we focus on the embedded approach and consider how this might be achieved.

So what AI techniques can be adopted, given the inherent limitations of mobile devices? Various techniques have been demonstrated in laboratory conditions but one paradigm has been demonstrated to be computationally tractable on mobile devices: intelligent agents. As well as forming the basis of mobile intelligent information's systems, a number of toolkits have been made available under open source licensing conditions thus allowing software engineers access to mature platforms at minimum cost. Before briefly considering some of these options, it is useful to reflect on the intelligent agent paradigm.

THE INTELLIGENT AGENT PARADIGM

Research in intelligent agents has been ongoing since the 1970s. Unfortunately, the term agent has been interpreted in a number of ways thereby leading to some confusion over what the term actually means. More precisely, the characteristics that an arbitrary piece of software should possess before applying the term agent to it are debatable. In essence, an agent may be regarded as a computational entity that can act on behalf of an end-user, another agent or some other software artefact. Agents possess a number of attributes that distinguish them from other software entities. These include amongst others:

- **Autonomy**: The ability to act independently and without direct intervention from another entity, either human or software-related
- **Proactivity**: The ability to opportunistically initiate activities that further the objectives of the agent
- **Reactivity**: The ability to respond to events perceived in the agent's environment;
- **Mobility**: The ability to migrate to different nodes of a network as the need to fulfill its objectives dictates; and
- **Social ability**: The ability to communicate with other agents using a shared language and ontology leading to shared or collaborative efforts to achieve individual and shared objectives.

To what extent an agent possesses or utilizes each of those attributes is at the discretion of the designer. For clarity purposes, it is useful to consider agents as existing on a scale. At the lower end are so-called reactive agents. Such agents act in a stimulus-response manner, and a typical usage scenario might involve the agent monitoring for user interaction and reacting to it. Such agents are generally classified as weak agents (Wooldridge & Jennings, 1995). At the other end of the scale are so-called strong agents. Such agents maintain a sophisticated model of their environment, a list of goals or objectives, and plans detailing how to achieve these objectives. Such agents support rational reasoning in a collaborative context and are usually realized...
as multi-agent systems (MAS). This strong notion of agenthood is synonymous with the view maintained by the AI community.

One popular interpretation of the strong notion of agency is that of the belief-desire-intention (BDI) paradigm (Rao & Geogeff, 1995). This is an intuitive and computationally tractable interpretation of the strong agency stance. To summarize: beliefs represent what the agent knows about its environment. Note that the term environment can have diverse meanings here and may not just relate to the physical environment. Desires represent the objectives of the agent, and implicitly the raison d’être for the application. However, at any moment in time, an agent may be only capable of fulfilling some of its desires, if even that. These desires are then formulated as intentions and the agent proceeds to fulfill these intentions. The cycle of updating its model of the environment, identifying desires that can be fulfilled, and realizing these intentions is then repeated for the duration of the agent’s lifecycle (Figure 2).

When should agents be considered for realizing a software solution? Opinion on this is varied. If the solution can be modeled as a series of dynamic interacting components, then agents may well offer a viable solution. However, many see agents as being particular useful in situations that are inherently complex and dynamic as their native capabilities equip them for handling the myriad of situations that may arise. Naturally, there are many situations that fulfill the criteria but, for the purposes of this discussion, it can be easily seen that the mobile computing domain offers significant opportunities for harvesting the characteristics of intelligent agents.

**Intelligent Agents for Mobile Computing**

As the capability of mobile devices grew, researchers in the intelligent agent community became aware of the feasibility of deploying agents on such devices, and perceived mobile computing as a potentially fertile area for the intelligent agent paradigm. A common approach was to extend the functionality of existing and well-documented MAS environments such that they could operate on mobile devices. It was not necessary to port the entire environment on to the device; it was just necessary to develop an optimized runtime engine for interpreting the agent logic. In this way, the MAS ethos is persevered and such an approach subscribes to the distributed AI approach alluded to previously. A further benefit was that existing agent-oriented software engineering

---

*Figure 2. Architecture of a BDI agent*
(AOSE) methodologies could be used. In the case of testing, various toolkits have been released by the telecommunications manufacturers that facilitate the testing of mobile applications. A prudent approach is of course to test the application at various stages during its development on actual physical devices, as this will give a more accurate indication of performance, the look and feel (L&F) of the application and so on. For a perspective on deploying agents on mobile devices, the interested reader should consult Carabelea and Boissier (2003).

While a number of environments may be found in the literature for running agents on mobile devices, the following toolkits form a useful basis for initial consideration:

1. **LEAP (Lightweight Extensible Agent Platform)** (Bergenti, Poggi, Burg, et al., 2001) is an extension of the well-documented JADE platform (Bellifemine, Caire, Poggi et al., 2003). It is FIPA (http://www.fipa.org/) compliant and capable of operating on both mobile and fixed devices.

2. **MicroFIPA-OS** (Laukkanen, Tarkoma & Leinonen, 2001) is a minimized footprint of the FIPA-OS agent toolkit (Tarkoma & Laukkanen, 2002). The original FIPA-OS was designed for PCs and incorporated a number of features that did not scale down to mobile devices. Hence, MicroFIPA-OS minimizes object creation, reduces computational overhead and optimizes the use of threads and other resource pools.

3. **AFME (Agent Factory Micro Edition)** (Muldoon, O’Hare, Collier & O’Grady, 2006) is derived from Agent Factory (Collier, O’Hare, Lowen, & Rooney, 2003), a framework for the fabrication and deployment of agents that broadly conform to the BDI agent model. It has been specifically designed for operation on cellular phones and such categories of devices.

4. **JACK** is, in contrast to the three previous frameworks, a commercial product from the Agent Oriented Software Group (http://www.agent-software.com). It comes with a sophisticated development environment, and like AFME, conforms to the BDI agent model.

A detailed description of each of these systems is beyond the scope of this discussion. However, the interested reader is referred to (O’Hare, O’Grady, Muldoon & Bradley, 2006) for a more advanced treatment of the toolkits and other associated issues.

**FUTURE TRENDS**

As mobile devices proliferate, and each generation surpasses its predecessor in terms of raw computational capacity and supported features, the potential for incorporating additional AI techniques will increase. In a similar vein, new niche and specialized markets for mobile services will appear. If a more holistic approach is taken towards mobile computing, it can be seen that developments in sensor technologies, fundamental to the ubiquitous and pervasive vision, will follow a similar trajectory. Indeed, the possibility of deploying intelligent agents on sensors is being actively investigated in widespread expectation that the next generation of sensors will incorporate processors of a similar capability to the current range of PDAs. Such a development is essential if the AmI vision to reach fruition.

As the possibility of incorporation of ever more sophisticated AI techniques increases, the potential for extending and refining the adaptability and IUI constructs for the support of mobile users increase. Indeed, adaptability may reach its fulfillment through the incorporation of autonomic computing precepts (Kephart & Chess, 2003). Self-configuring, self-healing, self-optimizing
and self-protecting are the key attributes of an autonomic system, and it can be seen that incorporation of AI techniques may make the realization of these characteristics more attainable.

Finally, the practical issues of engineering mobile AI solutions must be considered. Mobile computing poses significant challenges to the traditional software engineering process, and the broad issue of how best to design for mobile services still needs to be resolved. The situation is exacerbated when AI technologies are included. However, it may be envisaged that as experience and knowledge of the mobile computing domain deepens and matures, new methodologies and best practice principles will emerge.

CONCLUSION

Mobile computing scenarios are diverse and numerous, and give rise to numerous challenges that must be overcome if the end-user experience is to be a satisfactory one. IUIs offers one viable approach that software designers can adopt in their efforts to make their systems more usable in what is frequently a hostile environment. However, the pragmatic issue of realizing mobile applications that incorporate intelligent techniques is of critical importance and gives rise to significant technical and design obstacles.

In this chapter, the broad issue of realizing an intelligent solution was examined in some detail. At present, the intelligent agent paradigm offers an increasingly viable proposition for those designers who wish to include intelligent techniques in their designs. To illustrate the issues involved, the intelligent agent paradigm was discussed in some detail.

As mobile developments continue unabated, the demand for increasingly sophisticated applications and services will likewise increase. Meeting this demand will pose new challenges for software and HCI professionals. A prudent and selective adoption of intelligent techniques may well offer a practical approach to the effective realization of a new generation of mobile services.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Science Foundation Ireland (SFI) under Grant No. 03/IN.3/1361.

REFERENCES


Intelligent User Interfaces for Mobile Computing


**KEY TERMS**

**Ambient Intelligence**: (AmI) was conceived by the Information Society Technologies Advisory Group (ISTAG) as a means of facilitating intuitive interaction between people and ubiquitous computing environments. A key enabler of the AmI concept is the intelligent user interface.

**BDI Architecture**: The Belief-Desire-Intention (BDI) architecture is an example of a sophisticated reasoning model based on mental constructs that can be used by intelligent agents. It allows the modeling of agents' behaviors in an intuitive manner that complements the human intellect.

**Context**: Context-aware computing considers various pertinent aspects of the end-user's situation when delivering a service. These aspects, or contextual elements, are determined during invocation of the service and may include user profile, for example language, age, and so on. Spatial contextual elements, namely location and orientation, may also be considered.

**Intelligent Agent**: Agents are software entities that encapsulate a number of attributes including autonomy, mobility, sociability, reactivity and proactivity amongst others. Agents may be reactive, deliberative or hybrid. Implicit in the agent construct is the requirement for a sophisticated reasoning ability, a classic example being agents modeled on the BDI architecture.

**Intelligent User Interface**: Harnesses various techniques from artificial intelligence to adapt and configure the interface to an application such that the end-user's experience is more satisfactory.

**Mobile Computing**: A computer usage paradigm where end-users access applications and services in diverse scenarios, while mobile. Mobile telephony is a popular realization of this paradigm, but wearable computing and telematic applications could also be considered as realistic interpretations of mobile computing.

**Multi-Agent System**: A suite of intelligent agents, seeking to solve some problem beyond their individual capabilities, come together to form a multi-agent system (MAS). These agents collaborate to fulfill individual and shared objectives.

**Ubiquitous Computing**: Conceived in the early 1990s, ubiquitous computing envisages a world of embedded devices, where computing artefacts are embedded in the physical environment and accessed in a transparent manner.

*This work was previously published in the Handbook of Research on User Interface Design and Evaluation for Mobile Technology, edited by J. Lumsden, pp. 318-329, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*
Chapter 8.7
Intelligent User Interfaces for Ubiquitous Computing

Rainer Malaka
Bremen University, Bremen

ABSTRACT

Designing user interfaces for ubiquitous computing applications is a challenging task. In this chapter we discuss how to build intelligent interfaces. The foundations are usability criteria that are valid for all computer products. There are a number of established methods for the design process that can help to meet these goals. In particular participatory and iterative so-called human centered approaches are important for interfaces in ubiquitous computing. The question on how to make interfaces more intelligent is not trivial and there are multiple approaches to enhance either the intelligence of the system or that of the user. Novel interface approaches follow the idea of embodied interaction and put particular emphasis on the situated use of a system and the mental models humans develop in their real-world environment.

User interfaces for computational devices can be challenging for both their users and their designers. Even such simple things as VCRs or TV sets feature interfaces that many people find too difficult to understand. Reviews and tests of consumer electronic devices very often rank bad usability even higher than technical aspects and the originally intended main function of the devices or features. Moreover, for most modern appliances there is not much technical difference in their core functions. For instance TV sets differ less in quality of display and sound and more in the way the user interacts with the device. This already shows why user interface design is crucial for any successful product. However, we want to extend the question of user interface design in two directions: the user interface should become more intelligent and adaptive and we want more suitable interfaces for ubiquitous computing scenarios.

The first aspect seems to be clear at first sight: intelligent user interfaces are just what we want and nobody will neglect the need for smart, clever, and intelligent technology. But it becomes more difficult if we strip away the buzzwords and dig a bit deeper into the question of what an intelligent user interface actually should do and how
Intelligent User Interfaces for Ubiquitous Computing

it would differ from an ordinary interface. Would the standard interface then be a stupid one?

The second aspect introduces a new level of complexity: an interface is by definition a clear boundary between two entities. A user interface resides between human and machine; other interfaces mediate, for instance, between networks and computers. In ubiquitous computing we have the problem that there might not be a clear boundary any more. Computers are no longer visible and, in the end, they can disappear from the user’s conscious perception. We will, therefore, face the challenge of building an interface for something that is rather shapeless.

In the following, we will go into more detail through these questions and will introduce some general approaches for designing user interfaces. We will see that we can learn from good interface design for other—classical—devices, and that we can apply many of those user interface design principles for ubiquitous computing as well. A central aspect will be the design process that helps to find the right sequence of steps in building a good user interface. After discussing these general aspects of user interface design, we will focus on the specific needs for ubiquitous computing scenarios and finally on how to build intelligent user interfaces—or to be less euphemistic: to avoid stupid interfaces.

BUILDING GOOD USER INTERFACES

The design of a good user interface is an art, which has been ignored for a long time in the information and communication technology (ICT) business. Many software developers just implemented whatever they found useful for themselves and assumed it would also be beneficial for the respective users. However, most users are not software developers and their way of interacting with technology is very different. Sometimes, the result is technology that is highly functional and useful for a small group of people, namely the developers of the system, and highly inefficient, frustrating or even unusable for most other people. Some of the highlights of this dilemma can be found in the communication with the user when something goes wrong: An error message notifying the user: “an error occurred, code 127” might be of some use for the developer and help in his efforts in debugging the system, but a user will hardly be able to understand what went wrong.

Today usability plays a much bigger role and many systems (including computer systems) are now designed with more care for easy and safe usage. On the one hand this is due to legal constraints demanding accessibility, but also due to the fact that many systems do not differ so much in their technical details and vendors have to diversify their products solely in terms of their “look and feel.” We now have a wealth of methods, tools, and guidelines, which all help to develop a good user interface (Dix et al., 1998; Mayhew, 1999). However, there is not one single recipe whose application guarantees 100% success. The essence of usability engineering is to work iteratively in order to achieve the goal of better usability. Let us briefly go through these steps and summarize some of the most important issues of usability engineering. For more detailed information, a number of textbooks and research articles can be consulted (Dix et al., 1998; Nielsen, 1993; Shneiderman, 1997).

The first question of usability engineering is the question of what goals we actually want to achieve. The typical list of usability goals contains at least the following five (ISO 9241, 2006):

- **Safety and Security:** Good design should not harm users or other people affected by the use of a product. It should also help to avoid errors made by humans in using the system.
- **Effectiveness:** A good user interface supports a user in solving a task effectively,
that is, all aspects of a task can be actually handled.

- **Efficiency and Functionality:** A well-designed and usable system should allow for quick and timely work.

- **Joy and Fun:** How enjoyable is it to work (or play) with the system? Is it fun or is it a pain to interact with it?

- **Ease of Learning and Memorizing:** How fast can new users interact with the system and will they remember what they learned?

This list, of course, is not exhaustive and not all aspects can be fulfilled to the same (high) degree, which is to say that there are classic trade-offs. Some aspects, therefore, might even be in conflict with others and it is important to identify such conflicts and to decide which aspect to optimize and to what extent. For instance, when designing an interactive game, joy and fun might be more important and effectiveness is less important. In contrast, a system for firemen has to be more efficient and can be less fun. Another typical trade-off exists between the need for efficient work and for training. One solution can be to provide two modes: an expert mode and a novice mode.

As a general rule, all efforts and goals of usability should be measurable in quantitative or qualitative ways. And since most usability criteria depend on the actual use of a system, there is a need to involve users in the design process. Of course, many human factors have been studied and psychologists have theories about how people can perceive information and how they can—in principle—react. But, in order to actually find out if the goals are met, one must try things out with actual users. And the more unknown your application terrain is, the more involvement of users is required, which is of particular importance for ubiquitous computing because there is not yet a large set of experience, studies, and guidelines at hand.

The design process that involves users has been named *human-centered design* (ISO 13407, 1999). Its principle is to develop an application iteratively with evaluations in every cycle. Human-centered design also is regarded as the best approach when design goals are hard to formalize in technical terms.

There have been multiple approaches for system design processes that involve the users. Their roots are in the participatory design idea from Scandinavia that involves workers in the definition and design of their working environment (Olson & Blake, 1981). In contrast to the classical waterfall model in systems engineering (Royce, 1970) that segments the design process into a linear order of clearly separable steps, these models iterate and involve users and evaluations in each cycle. A number of models have been proposed replacing the waterfall scheme by cycles or stars, that is, the design process is open and decisions can be revised depending on user feedback during development (Gould et al., 1991; Hartson & Hix, 1989; Hix & Hartson, 1993). Since many usability goals are not well defined and cannot be formally defined beforehand, these models allow for a continuous evolution of the usability of the system (Figure 1).

The design steps in these models are the following:

- **Definition of the Context:** As a first step, designers should consider the context of their envisioned product. This includes defining the way the system will be used, if it will be used for life-critical or fun purposes, and in home or in office environments, as well as the market situation. The latter is important because it tells something about expectations of users and about who is going to buy the product. In general, not only the target users are involved in deciding about the success (i.e., sales) of a product. Decisions can be made by managers of the users and they can
influence third parties such as the customers or clients of the users.

- **Description of the Users**: Based on the context definition, each group of directly or indirectly affected users must be carefully analyzed. Their physical and cognitive abilities and their cultural and social background may affect the way they interact with the system. Special needs may play a role. Accessibility has become important for IT systems and is demanded by many legal constraints, in particular in working environments.

- **Task Analysis**: Multiple techniques help to derive a rather formal description of the task users want to solve from informal interviews and observations. Most importantly, designers should find out how users actually solve their task currently (not how they think they do it) and how they make use of tools at hand, how they communicate and how their context influences the course of activities.

- **Requirements/Specification**: This step would have been the first step of the classical software development process. For user-centered design, it is now based on a better understanding of the users, their context and their tasks. Moreover, the specifications can be changed in each iteration, when a better understanding of the system could be gained through evaluations.
Intelligent User Interfaces for Ubiquitous Computing

- **Conceptual Design/Formal Design:** The requirements and specifications are translated into system components.
- **Prototyping:** Instead of “doing it right the first time” we iteratively build prototypes of the system. A prototype can be a very simple design sketch or an almost complete and working system depending on the stage of iterations in the design process.
- **Evaluations:** Evaluations are essential for assessing the progress of the design process and for deriving a better understanding of the tasks, the requirements and thus a better specification and implementation of the system (prototype).
- **Implementation, Tests, Maintenance:** Whenever the iterations have got to a stage where the prototype sufficiently fulfills the design goals, the final prototype (product) can be implemented. Of course tests and maintenance are as important as in classical system engineering. Moreover, they can help to further improve the system and in particular user feedback after deployment can be used for defining new development cycles.

These design steps are the building blocks for good user interface design. They are very generic and they are valid for basically every interactive system. Iterative development, however, is inevitable for the design of human-computer interaction in ubiquitous computing as we enter a domain of interactive systems, where we cannot derive system requirements from interaction goals without user involvement. This is mainly due to the fact that interaction in ubiquitous computing aims at intuitively usable pervasive IT systems that assist users in their real-world endeavors. Without taking these aspects into account, these systems are subject to failure. Many ubiquitous computing prototypes are completely technology-driven. Their developers focus on smart new gadgets, networks and infrastructure but they do not focus their design efforts on their users. Just for the sake of plausibility, some usage scenarios and users are added to the design. Such systems will not leave the research labs and they will fail to find their market.

**EVALUATIONS, ALTERNATIVES, AND PROTOTYPES**

The importance of iterations in designing intelligent user interfaces for ubiquitous computing has now been emphasized. However, how should that bootstrapping actually happen? Where to start, how to proceed and when to stop? If we need a full-fledged prototype in each iteration along with an evaluation with a high number of users, the costs for developing a better ubiquitous computing application will rapidly explode. Fortunately, things can be done more efficiently and some techniques help to manage the process.

**Where to Start?**

In order to get a first impression of how to build a system that actually meets the usability goals, for example, being an understandable and enjoyable assistant for some user task, we do not need any system but can make up a fake system without bothering with how to build a real one. A number of methods can be used (Dix et al., 1998; Shneiderman, 1997):

**Design Sketches**

Instead of actually building something that looks like a real system, users or usability experts can evaluate early design ideas. First sketches on paper or on a blackboard can already give an impression of the designer’s ideas, and feedback can already help to avoid basic mistakes. Moreover, the discussion can facilitate the mutual understanding of the users’ world and the prospective system.
Wizard of Oz Experiments

If, however, the users should already get an impression of how the interaction with the system might look, a system can also be simulated. A human operator remote controls all functions of the environment and the test users are told they are already interacting with the system. This technique has been proven to be extremely fruitful for systems that need data on the interaction in advance. For instance for systems that are language controlled, Wizard of Oz experiments can be used to collect utterances and language patterns that help to build speech recognizers and grammars for the real system.

Mock-Ups

A mock-up is a model of the system that already exposes the “look and feel” but does not yet include real functionality of the intended system. Early mock-ups for graphical interfaces can, for instance, consist of a PowerPoint walkthrough through a system or some Web sites emulating a system.

Prototypes

In contrast to the mock-up, the prototypes include actual functionalities of the target system. They may iteratively evolve to the final system.

Since many applications for ubiquitous computing scenarios are embedded into real-world tasks and many of them are also affected by or affect other objects in the users’ surroundings, Wizard of Oz experiments are a cheap and very beneficial first step in system design. They can help to understand how people would interact in an environment that is enhanced by ubiquitous computing technology. Moreover, the designers get data that help to design interaction with the system. For most cases of more natural interaction such as speech or gesture, such data is necessary anyway because the recognizers need training data.

How to Proceed?

Evaluation is the core of the above-mentioned “star-model.” Depending on the maturity of the design, the budget and the nature of the system, a great variety of evaluation techniques can be used. Evaluation methods can be classified according to the following dimensions:

- **Qualitative vs. Quantitative Methods:** In qualitative methods, feedback in form of comments, impressions and subjective ratings is collected in interviews or questionnaires. Quantitative methods measure parameters such as error rates, task completion times or movements of users in order to estimate the quality and efficiency of an interface.

- **Studies in the Field or in the Lab:** Field studies are conducted under realistic conditions where the systems are actually used, for example, in the office or home of the users. They usually need more effort than studies in the lab under simulated conditions, but they yield more realistic results.

- **User Tests or Expert Evaluations:** User studies involve real test users. They are more expensive than expert evaluations where a few experts judge the system by their experience on user behavior and the application domain. There are many well-known techniques for both—such as cognitive walkthrough, discount evaluation, thinking aloud—and in some cases even combinations may be useful.

- **System State (Sketch, Mock-Up, Prototype, …):** As discussed above, in early evaluations, a system does not necessarily have to be fully functional but can rather be a sketch or a mock-up.
It is beyond the scope of this chapter to go into all details of evaluation techniques. We will focus rather on the most important aspects for ubiquitous computing interfaces.

Even though evaluation is crucial for the design of good interfaces, it should be noted that evaluation techniques do not solve all problems and can even be misleading. One of the main problems of evaluations is that they are always limited snapshot observations restricted in the time of usage and the complexity of the context. This is important to note, in particular, for ubiquitous computing systems interfaces. Take, for instance, the famous ubiquitous computing scenario of an intelligent refrigerator that keeps track of its contents and can alert a user when she is running out of milk. In an evaluation setting one could look at users while they are at home or while they are in a supermarket and one could measure how they react to notifications of the system. A questionnaire reveals if the users like the system and would like to buy it when it comes on to the market. In a realistic setting, a study would observe some 10 to 20 users each over a time span of one to two hours of interaction. All would be in the same representative supermarket and in some model kitchen. The results would be definitely interesting and the study would even go beyond many other evaluations of similar systems. However, it is too limited for multiple reasons:

- **No Long-Term Observation:** Since users would interact with such a ubiquitous computing system not only for a few hours but rather over months or years, the short interaction of a novice user does not reveal much about the user’s future interaction.

- **Limited Frame of Context:** In order to gain comparable results, all users are set to the same or a similar context. In everyday situations, however, contexts may differ a great deal and users show a much higher degree of variation in their behavior.

- **Additional Tasks, People, and Devices:** As with most ubiquitous computing applications, users may not be focused on just one task but may be doing many other things concurrently. They could have other devices with them or be interacting with their colleagues or family members.

These limitations of evaluation results make some of them questionable. However, by using a good and careful evaluation design, some aspects can be counterbalanced. Moreover, keeping the limitations in mind may help to focus on the right questions and avoid overstating the results. And finally: even when evaluations only shed limited light on the usability of a system, this is much better than working in complete darkness without evaluations.

As a rule of thumb, it should be noted that evaluations for ubiquitous computing interfaces should be made as realistic as possible. Thus field studies would be better than lab conditions. Moreover, the designers should have a clear understanding of what they want to achieve with their system in order to know what they want to prove using evaluations.

### When to Stop?

The development cycle should not be an endless loop. In general, the (re-)design-prototype-evaluation cycle can go on forever leading to a continuous increase of usability. In practice, either the number of cycles is fixed beforehand or certain measures define when the loop has to be stopped and the final design is achieved. Typically these measures would quantify the usability goals listed at the beginning of this chapter. Such a goal could be “95% of the test users rate the system as very convenient” or “the task completion rate within 30 minutes is 98%.” In some cases the stop-criterion is not bound to usability but to other measures.
such as “we are out of budget” or “the deadline is next week.”

**SPECIFIC CHALLENGES OF USER INTERFACES FOR UBQUITOUS COMPUTING**

So far we have learned about how to design a good user interface. The principles we discussed are rather generic and they apply—of course—for designing intelligent user interfaces for ubiquitous computing, but they are also valid for other user interfaces such as Web interfaces or interfaces of desktop applications. The general process of human-centered design could even be applied to non-IT products such as cars, coffee machines and other objects of our daily life. It is a matter of fact that we have got to such a generic process. On the one hand, good usability is a property that is generic and the design process is fairly similar in multiple domains. On the other hand, ubiquitous computing is about integrating things into the objects of our normal life. Thus usability has, owing to the very nature of ubiquitous computing, got something to do with the usability of everyday things.

Since the early days of ubiquitous computing, usability has been in its focus. Mark Weiser’s idea of ubiquitous computing encompasses “invisible” interfaces that are so naturally usable that they literally become invisible for the user’s conscious perception (Weiser 1999a, b). This notion goes back to the German philosophers Georg Gadamer and Martin Heidegger, who call such interaction with things that we use without conscious awareness things that are “ready-to-hand” or at our “horizon.” In this phenomenologist view, the meaning of the things is actually derived from our interaction with them. Such a view on interactive artifacts has become popular in ubiquitous computing and is closely related to the notion of embodiment (Dourish, 2001). This is a fundamental shift from the classical positivist approach in computer science, that is, modeling the real world in simplistic formal computer programs, to an embodied approach that takes the user in the real world into account. This is relevant for ubiquitous computing for multiple reasons. On the one hand, ubiquitous computing applications are to be used in complex real-world settings and their meaning (for the user) will, in fact, only evolve in the course of action. Additionally, if things should become natural extensions of our physical abilities, they must be designed such that they do not need conscious interference from their users.

Given this notion of being “invisible” we can see that this does not necessarily mean “not there,” but rather present without conscious interaction. The most basic examples for such physical objects are our body parts. We do not have to think consciously about what we do with our arms, but we just do the things we want. When we leave our house, we do not have to remember: “let’s take the arm with us, we might need it today.” It is there and ready for immediate use. When we throw a ball, we just throw it and we do not think and plan how to make our hand grasp the ball and our arm swing around in order to accelerate the ball. In this sense, our arm is invisible but also very present. Thus if we speak of ubiquitous computing interfaces that are “invisible” or computers that are “disappearing,” we actually speak of things that are present and “ready-to-hand.” However, the artifacts we interact with might not be consciously realized as computers.

A good example of such a “ubiquitous” technology is present in our homes already: electrical light. Whenever we enter a room that is dark, we just find a switch with our hands next to the door and the light goes on. Without thinking we turn on the light. We do not think of cables that conduct electrons. We do not have to consider how the light bulb works or how they generate electricity at the power plant.

We have a very simplistic model of how the thing works and it is internalized to such a de-
Intelligent User Interfaces for Ubiquitous Computing

gree that we do not have to think about it when we enter a room. These “mental models” of how things work play an important role in designing good user interfaces as well as in designing other everyday things (Norman, 1998). Donald Norman emphasizes that a good design is about providing good mappings (Figure 2):

- The design model must be mapped to the system image.
- Users must be able to map their understanding (mental model) to the system.
- The system must allow the user to map its image to the user’s model.

The question is now: how can a system image support the appropriate user’s mental model? The answer—with our notion of embodiment in mind—must bring the meaning of things into the things themselves and thus a user can derive the meaning of something from the interaction with it or from its mere appearance that may signal some properties indicating how to use it. Such properties have been named affordances (Norman, 1998). The idea of affordances is to bring knowledge into the world instead of having it in mind. Many highly usable things that surround us just let us know by their physical appearance how we can use them. A chair for instance does not need a label or instructions on how to sit on it. We just see and know it is a chair and we know what to do with it.

Similarly, affordances have been defined as virtual affordances for computer interfaces and many metaphors on our computer screens signal functionalities, for example, mouse pointers and scrollbars. With the advent of ubiquitous computing, the term affordance becomes again more literally a property attached to the physical properties of things. Many ubiquitous computing objects include tactile interfaces or smart objects with physical and not just virtual properties.

There are a number of consequences arising from this perspective of embodied interaction for ubiquitous computing:

Figure 2. Mappings of design model, mental model and system images (Norman, 1998)
• **Support Mental Models:** Humans use mental models to understand and to predict how things react to their actions. The system image should support such mental models and make it easy to understand it.

• **Respect Cognitive Economy:** Humans re-use their mental models. If there are well-established mental models for similar things, then they can be a good basis for an easy understanding of a new artifact.

• **Make Things Visible and Transparent:** In order to understand the state of an object it should be obvious what is going on. For instance, a container can indicate if it is loaded or not.

• **Design for Errors:** Mappings between the user's model and the system sometimes fail. Most “human errors” are, in fact, mapping errors. Therefore, systems must assist users in finding a solution for their task even if something has gone wrong. There are a number of techniques for doing so, for example, allowing undo-actions or sanity checks on user inputs.

• **Internal and External Consistency:** Things within an application should work consistently. For instance, pushing a red button always means, “stop.” External consistency refers to expectations users may have from usage of other applications. If we add some ubiquitous computing technology to a cup and turn it into a smart cup, a user will still expect the cup to work as a cup.

With these guidelines and the general design process considerations we are already well prepared for building very good interfaces for ubiquitous computing applications. However, there are a number of further practical considerations and human factors that play a role for ubiquitous computing user interfaces. Some of these issues are related to the very nature of these applications being “ubiquitous” and some are more related to technical problems in mobile and ubiquitous scenarios. We will briefly highlight some of these aspects. Due to the broad spectrum of possible applications, we cannot go into details of all possible factors.

### Human Factors for Ubiquitous Computing

In classical human-computer interaction, we have a well-defined setting. In ubiquitous computing, we do not know where the users are, what tasks they are doing currently, which other persons may be around. This makes it very hard to account for some human factors that can greatly influence the interaction. Depending on time, concurrent tasks, and so forth, the user’s cognitive load, stress level, patience, and mood may vary extremely. Thus an interface can, in one situation, be well suited and in another situation the user is either bored or overloaded.

Another problem lies in spatial and temporal constraints. In many ubiquitous computing applications, location and time play a crucial role. Users need the right information at the right time and place. In a system that helps a user to navigate her vehicle through a city, the information “turn right” only makes sense at a very well defined point in space and time. An information delay is not acceptable. Even though space and time are the most prominent context factors in systems today, other context factors may also play a big role (cf., the chapter “Context Models and Context Awareness”). An interface can adapt to such context factors and take into account what is going on. In particular, the user might not have the focus of attention on the system but rather might be busy doing something else. But not only user-driven activities can distract the user; other people and events are not the exception but the normal case in many ubiquitous computing scenarios. This has a huge effect on the interface and dialog design. While in desktop applications, the designer can
assume that the user is looking at the screen and a system message is (in most cases) likely to be read by the user, in ubiquitous computing we must reckon with many signals from the system being ignored by the user.

The interfaces can try to take the users’ tasks into account and thus adapt their strategy to reach the user’s attention. For example, when the user is driving a car, the system might interact in a different way than when the user is in a business meeting. However, when the system is literally ubiquitous, the number of tasks and situations the user might be in can be endless and it is not feasible to model each and every situation. The system interface might then instead be adaptable to a few distinct modes of interaction.

**Who is in Charge?**

As we mention adaptation and adaptivity, we get to a point where the system behaves differently in different situations. This can be a perfect thing and can significantly increase the ease of use. A mobile phone, for instance, that automatically adapts to the environment and stays silent in a business meeting, but rings in other situations is rather practical. However, the trade-off is a reduced predictability and, as discussed above, many usability goals can be in conflict with each other. The developers and (hopefully) the users have to decide which goal is more important. It is important to know about these conflicts and to decide explicitly how to deal with them.

Typically, usability goals in ubiquitous computing that come into conflict with others are:

- **Controllability:** Is it the system or the user who controls the situation?
- **Support of Mental Models:** How can a user still understand a very complex system?
- **Predictability:** Humans want to be able to predict the outcome of their actions. If a system is too adaptive and autonomous, users get lost.
- **Transparency:** If the system adapts to all sort of context factors, its state becomes less transparent.
- **Learn Ability:** A system that learns and behaves differently in new situations can be hard to understand.

The designers have to decide to what degree they want to achieve which level in each of these dimensions and how other aspects such as autonomy or adaptivity may affect them. In general, there are no rules or guidelines that can give clear directions. While in many other IT domains, such as Web systems, some established standards may set the stage and good guidelines exist, the designer of a ubiquitous computing system will have to derive his own solution on the basis of the goals he wants to achieve. The only way to prove that the solution actually fits these goals, are, in turn, evaluations. Therefore, a user-centered design approach is the only way to design ubiquitous computing systems that incorporate good user interfaces.

**INTELLIGENT AND DUMB INTERFACES FOR UBQUITOUS COMPUTING**

In the last part of this chapter we want to focus on intelligent user interfaces. The term “intelligent user interface” has been debated for a while and it is not so clear what it means and if at all intelligent interfaces are something beneficial. But even the term intelligence is not well defined and has been used (or misused) in multiple ways. Before going into technical details we should, thus, first discuss what the term means and then see some techniques that are used for realizing them. We will finish with a discussion on how much intelligence a good interface actually needs.
What is an Intelligent User Interface?

So far we have presented a number of techniques for building good interfaces. We also saw how the view of embodied interaction can be used as a paradigm for ubiquitous computing. In general, a technical solution can be called “intelligent” for two reasons: (1) there is some built-in intelligent computation that solves some otherwise unsolvable problem; (2) using the system, a user can solve an otherwise unsolvable problem, even though the system itself does not actually do anything intelligent. Suppose that calculating the logarithm of a number is a hard problem for a human, then a calculator is a good example for case (1) and an abacus would be an example for (2). The calculator solves the problem for the human and the abacus empowers the user to solve the problem on her own.

The classical approach of artificial intelligence (AI) is a rationalist one. According to this approach, a system should model the knowledge that human experts have and thus emulate human intelligence. In this sense, the “intelligence” moves from the user to the system (Figure 3a, 3b). This approach is valid for many cases, for example, if expert knowledge is rare and non-experts should also be able to work with a system. As discussed above, the embodied interaction view would rather try to make the interaction more intelligent (Figure 3c). This fits too many new trends in AI where embodied intelligence is viewed as a property that emerges from the interaction of an intelligent agent with the environment. In this view, even simple and lightweight agents can perform intelligent behavior without full reflective and conscious knowledge of the world. With respect to this definition, all of the above-mentioned material already describes how to build an intelligent interface. Because the processes for designing human-centered systems are just the right techniques for designing intelligent interac-

Figure 3. Multiple views on intelligent user interfaces
tive systems, we already defined to a great extent how to build intelligent user interfaces.

Instead of leaving all the intelligence to the system, the user or the interaction, we can also try to get the best of all worlds and combine these techniques into a cooperative system, where both the system and the user cooperate with their knowledge on solving some tasks supported by intelligent interaction techniques (Figure 3d).

As discussed above, we can make the system more intelligent by enhancing the system, the interaction or the user. Intelligent user interface techniques exist for all three aspects. We will briefly list the key methods. Some details on them can be found in other chapters of this volume.

Techniques for Enhancing the System’s Intelligence

A huge number of AI techniques can be used to put more knowledge and reasoning into the system. Besides state-of-the-art IT methods such as databases, expert systems, heuristic search and planning, a number of more recent developments have attracted a good deal of interest by researchers and practitioners in the field:

- **World Knowledge and Ontologies**: Semantic technologies and formal models of world knowledge have had a great renaissance in the last couple of years. In context of the Semantic Web efforts, ontologies have been established as a standard method for capturing complex relations of objects and events in the world. Ontologies (cf., the chapter "Ontologies for Scalable Services-Based Ubiquitous Computing") can be successfully used in user interfaces in order to give the system a better understanding of the domain of an interaction. In particular for natural language interaction, ontologies provide resources for better understanding and reasoning.

- **User Adaptation**: User models and techniques of user adaptation allow for individualized interaction (cf., the chapter "Adapting to the User"). A number of methods allow for autonomous and user-driven customization of systems and they are widely used for intelligent user interfaces. In particular for ubiquitous computing, user adapted systems play a big role since these systems often have to support a great variety of use cases where a single standardized interface is not appropriate.

- **Context Adaptation**: Context plays a crucial role for ubiquitous computing (cf., the chapter "Context Models and Context Awareness"). As discussed already, context-dependent user interfaces can greatly enhance the usability of these systems. However, context can also be challenging because it can depend on a huge number of parameters and it is hard to formalize the meaning of contexts and to learn the relations between them autonomously.

- **Service Federation**: Integrating a variety of services and providing a single interface can be a significant step towards intelligent user interfaces. If users do not have to interact with all sorts of services separately, but can use a single portal, they can work much more efficiently with less cognitive load. However, service integration can be a hard problem, in particular when multiple services have to be integrated semantically that had not originally been designed to be integrated. An intelligent ubiquitous travel assistant could, for instance, integrate maps, events, travel, sights and weather information from different providers and offer the user an integrated trip plan.

Of course, many other techniques are available. In principle all advanced semantic and AI-based methods with relation to user interaction can help
to make systems smarter and better understand what a user might need and want using ubiquitous information sources.

**Techniques for More Intelligent Interaction**

As discussed above, the most important aspect of intelligent interaction is to provide good and working mappings of the user’s models of the world and the system’s model. These mappings depend highly on the semiotics, that is, the meaning and perception of the signs and signals established between user and the system. These can be both actively communicated codes in form of a language but also passive features of the artifact that signal the user affordances. Both aspects can be supported through intelligent methods that aim at more natural interaction such that the interaction takes place based on the premise of human communication rather than machine languages.

- **Multimodal Interaction:** Multimodal techniques make use of the human ability to combine multiple input and output modalities for a semantically rich, robust and efficient communication. In many ubiquitous computing systems, language, gestures, graphics, and text are combined to multimodal systems (cf., the chapters “Multimodal and Federated Interaction” and ”Multimodal Software Engineering”). Multimodality is on the one hand more natural, and, on the other hand, it also allows for more flexible adaptation in different usage situations.

- **Cross-media Adaptation:** Ubiquitous computing systems often use a number of different media, devices and channels for communicating with the user. A user can, in one situation, carry a PDA with a tiny display and, in another situation, interact with a wall-sized display in public or even use no display but just earphones. Intelligent interaction can support media transcoding that presents content on different media adapted to the situation.

- **Direct Interaction:** Humans are very good at multimodal communication, but for many tasks we are even better using direct interaction. It is, for instance, much easier to drive a car using a steering wheel than to tell the car to which degree it should steer to the left or to the right. For many activities, direct interaction is superior to other forms of human-computer interaction.

- **Embodied Conversational Agents:** Since humans are used to communicating with humans (and not with machines), anthropomorphic interfaces presenting animated characters can in some circumstances be very useful. In particular in entertainment systems, so-called avatars have become quite popular. However, there is also some debate about these interfaces and some people dislike this form of interaction.

Again, the list of techniques could be much longer. Here we just highlight some of the most important trends in the field. More ideas have been proposed and there will be more to come.

**Techniques for Amplifying the User’s Intelligence**

In principle, we can try to build better interface techniques, but we will not be able to change the user’s intelligence—leaving aside e-learning and tutorial systems that might explicitly have teaching or training purposes. But even if we do not affect the user’s intelligence, we can still do a lot about her chances to make use of it. In the scientific community there has been a controversy about the goal of intelligent user interfaces over the last couple of years on where to put how much intelligence (Figure 3). And even though it might be counterintuitive, an intelligent interface can sometimes be the one that leaves the intelligence on the part of the user rather than putting it into
Intelligent User Interfaces for Ubiquitous Computing

In the debate about these approaches, the slogan “intelligence amplification (IA) instead of artificial intelligence (AI)” was coined. The idea is that a really intelligent interface leaves intelligent decisions to the user and does not take away all intelligent work from the user by modeling it in the system. The question is: how can the system support users in acting intelligently? The answers have been given already when we discussed usability goals: A system that is easy to learn, where users have control and understand what is going on, and where mental models are applicable is more likely to let people act intelligently. In contrast, a system that only leaves minor steps to the user, that does not provide information about its states and how it got there and for which the users do not have appropriate mental models will in the long run bore its users and decrease their creativity and enthusiasm.

It should be noted that both type of systems can be packed with AI and be extremely smart or very dumb things. It is more the kind of interaction design that facilitates human intelligence or not.

HOW MUCH INTELLIGENCE?

Intelligent user interfaces for ubiquitous computing will be a necessary thing in the future. However, there are multiple competing views and philosophies. In general, three things could be intelligent: the user, the system or the way in which they interact. Most researchers focus on enhancing the system’s intelligence and the assumption is that this will lead to a better usability. This is often the case but not always. A different approach is to say that users are the most intelligent agents and their intelligence should be enhanced rather than replaced by artificial intelligence (IA instead of AI). In practice, however, we should do all together in order to make the interaction as easy and efficient as possible. But each decision should be made carefully keeping in mind that the overall goal of an intelligent user interface still should be defined by the usability goals. And like with all good things “less is sometimes more” and some simple things often are more enjoyable and easier to understand than highly complex and automated devices.

CONCLUSION

This chapter introduced aspects of designing user interfaces for ubiquitous computing in general and intelligent interfaces in particular. The basics for building intelligent interfaces are techniques for building good interfaces. Consequently, we first presented an up-to-date introduction to methods of human-centered design. A central aspect of this technique is to iteratively design systems with repeated evaluations and user feedback. This approach is especially important for ubiquitous computing systems, since they lack clear guidelines and decades of experience; and thus iterations are crucial in order to approach the desired design goals.

Obviously, many of these basic techniques are also valid for many other systems. However, ubiquitous computing introduces some more specific issues, such as a high variety of contexts, the lack of single dedicated interface devices and—by its very nature—ubiquitous interaction at any time and location. Therefore, ubiquitous computing interfaces must place even more emphasis on good mappings to mental models and provide good affordances. The view of embodied interaction gives us a good theoretical idea of the way we should think of and model interaction in such systems.

With these prerequisites, designers can build very good interfaces and can take many usability aspects into consideration. However, so far the “intelligence” of the interfaces has not been discussed. We did that in the last part of the chapter and presented a modern and sometimes contro-
versial view on intelligent user interfaces. There are different paradigms that may contradict each other. The main question can be formulated as: “AI or IA—artificial intelligence or intelligent amplification?” We discussed these design philosophies and presented some ideas on how to combine the best of both worlds. We also presented a number of current trends in the field that can be found in modern ubiquitous computing systems.

FUTURE RESEARCH DIRECTIONS

As the field of ubiquitous computing matures, its user interface techniques will also undergo an evolutionary process and some best practices will be established, making things much easier. We currently see this happening for Web applications where developers can choose from established interaction techniques that are well known to the users and guarantee efficient interaction.

However, ubiquitous computing might never reach that point since the ambition to support users in every situation at every time and place, which is the final goal of it, requires such rich interfaces that have to cope with the complexity of the users’ entire life. This might be good news for researchers in the field because they will stay busy searching for better and more intelligent interfaces.

The main challenges for future research will lie in the problem of extensibility and scalability of intelligent user interfaces. How could a system that has been designed for a user A in situation S be extended to support thousands of users in a hundred different situations?

REFERENCES


Intelligent User Interfaces for Ubiquitous Computing


ADDITIONAL READING


This work was previously published in the Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises, edited by M. Mühlhäuser and I. Gurevych, pp. 470-486, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 8.8

Vive la Différence:
The Cross-Culture Differences Within Us

David Gefen
Drexel University, USA

Nitza Geri
The Open University of Israel, Israel

Narasimha Paravastu
Central State University, USA

ABSTRACT

In the ITC cross-cultural literature, we often talk about the differences among peoples and how their respective culture and history may affect their adoption and preference usage patterns of ITC. However, do we really need to look that far to find such cross-cultural differences? Considering language is one of the major defining attributes of culture, this article takes a sociolinguistic approach to argue that there is also a cross-cultural aspect to ITC adoption within the same culture. Sociolinguists have claimed for years that, to a large extent, the communication between men and women, even within the supposedly same culture, has such characteristics because men and women communicate with different underlying social objectives and so their communication patterns are very different. This article examines this sociolinguistic perspective in the context of online courses. A key finding is that although the stage is set to smother cultural and gender differences if participants wish to do so through ITC, gender based cultural patterns still emerge. These differences were actually strong enough to allow us to significantly identify the gender of the student, despite the gender neutral context of the course discussions. Implications for ITC, in general, in view of this Vive la Différence, are discussed.

INTRODUCTION

One of the major manifestations of culture is language and the way it affects communications: who we prefer to talk to and the sum of the underlying objectives of the communication.
Communication is not a mere exchange of words. It is a social process and, as such, it is imbued with a social meaning of inclusion, exclusion, and social hierarchy. These cultural aspects are a prime aspect of cross-cultural research, including in the context of ICT information technology and communications, adoption, and usage patterns. But one need not look that far to find cross-cultural differences. They are here among us all of the time—that, at least, is the basic premise of sociolinguistics.

Sociolinguistics deals, among other things, with the way culture affects and determines communication. Most important in the context of this study is that culture is not only a manifestation of language and national heritage. Culture is also a matter of gender. Men and women communicate differently, and do so with different underlying social objectives. This is part of our evolutionary past (Brizendine, 2006), which also affects online collaboration (Kock & Hantula, 2005). Gender is so much a part of communication that in many languages, there are distinct rules in the language about how men and women should conjure the sentences they speak and their expected speech patterns. It is much more than superimposed linguistic gender segregation though. It is, at least in the view of sociolinguists, a matter of a cultural difference between men and women.

In general terms, men, according to sociolinguistics, communicate more with the objective of creating and preserving their social status, while women communicate more with the objective of creating rapport and social inclusion. Not surprisingly, the result of this is that communication across genders is often an exercise in cultural miscommunication (Brizendine, 2006; Tannen, 1994; Tannen, 1995). Indeed, when men communicate with each other, it is often on a basis of exchanging information, or as Tannen calls it “report talk,” while women do so to exchange emotions, or as Tannen calls it “rapport talk” (Tannen, 1994). The consequence of this is often communication that are gender segregated (Hannah & Murachver, 1999; Yates, 2001). Looking at this distinction in the context of virtual communities and supporting it, Gefen and Ridings (2005) commented that when men joined virtual communities composed of mostly male members, they did so with the declared objective of sharing information, while when they joined mixed gender virtual communities, it was more for emotional support. In contrast, women who joined mixed virtual communities did so for information exchange, but when they looked for emotional support, they too joined mostly female ones. Indeed, even in what should be gender and emotion neutral settings, women perceive more social presence in e-mail (Gefen & Straub, 1997) and e-commerce websites (Gefen, 2003).

The objective of this study is to examine whether the expected gender-related cultural differences in oral communication, predicted by sociolinguists regarding oral communications, hold true also in the explicitly created gender-neutral ITC environment of online courses, where the nature of the controlled course conversations make social dominance and rapport rather irrelevant. If these gender communication patterns hold true also in this scenario, then how much more so that such cross-cultural differences should hold true in other ITC induced environments. This is a crucial question because if true, then cross-cultural research in ITC should look not only across the border, but also within.

The data support the basic Vive la Différence proposition of the study even in the stoic context of online course discussions. Male students did prefer to respond to other male students and female ones to females, and men did show a more domineering attitude in their postings. Cross-cultural studies in ITC should consider gender as another dimension of culture.

Theory

The tendency of society toward being masculine or feminine is a central aspect of the cultural di-
mensions of peoples (Hofstede, 1980; Hofstede, Neuijen, Ohayv, & Sanders, 1990). However, is gender also an aspect within a culture? According to sociolinguistics, it is (Yates, 2001). In fact, the popular press sometimes even takes it a step further claiming, perhaps jokingly, that this gender difference might even take on celestial proportions (Gray, 1992). If this is so, then gender-related social behavior should come through even in the stoically enforced context of online courses, thus demonstrating the need to include gender as an aspect of culture even within a given national and linguistic culture, as also biology implies (Brizendine, 2006).

One of the major manifestations of culture is in language and communication. Communication carries with it not only information, but often also a very strong social message—a social message which is interpreted and sometimes also misinterpreted, within the cultural context of the speaker and listeners. Even the very way words are pronounced carries a cultural burden with it, making people identify or not with the speaker, based on the national or local culture implied in the accent (Deaux, 1984; Deschamps, 1982). Speaking in the accepted dialect can in fact make all the difference between whether people agree or disagree with a speaker, based almost purely on the manifestation of the presumed culture of the speaker (Abrams & Hogg, 1987). This additionally crucial social level of communication is a function not only of national and local culture but, according to sociolinguists (Tannen, 1994; Yates, 2001) and hormones (Brizendine, 2006), also of gender. Men and women may communicate in what, on a superficial level, may seem as the same language, but the social message behind the words and this message is interpreted quite differently by the average man and the average woman. This is because men and women, on average, imbue and insert different social nuances into the message and do so even in languages, such as English, where there are no linguistically gender enforced styles. These gender related nuances can be so manifest as to result in the equivalent of cross-culture miscommunication (Tannen, 1994).

Picking up on this idea, Gefen and Straub (1997) showed that women, across cultures, sense more social presence in work-related e-mails and that increased sense of social presence affect their perceptions of the usefulness and ease of use of the ITC and ultimately its usage. Expanding on this theme, Venkatesh and Morris (2000) showed that women are more affected by social norms in their adoption of ITC.

A salient example of this underlying social message in communication brought by Tannen (1994) is asking for directions. On the face of it, asking for direction is no more than just asking a stranger a question in what may seem a neutral environment. It could be regarded as information exchange and no more. However, this is not the case. Asking for directions also carries a social meaning. That is why men will often drive around for hours rather than ask for directions, while women will think nothing of it and do so without hesitation when they think they are lost. The reason for this, sociolinguists say, is that in asking for directions, men are subconsciously implying at least to themselves (certainly women are often surprised to hear this) that the other guy knows more than they do. The person being asked for directions certainly may know more, after all that is why they are being asked for directions, but it is admitting this that bothers men and makes them drive around for hours. Admitting someone else knows more than I do, to men, carries with it a social inferiority message of the other guy is better than I am in something. Men, unless aware of the stupidity of this underlying message, are loath to admit this supposedly social inferior standing. The same communication with women, however, carries no such subconscious implication. If anything, to women this creates a chance to engage with others, rapport, which they more willingly do than men (Tannen, 1994).²

This example, adapted from Tannen (1994), highlights the cultural social difference in com-
munication between men and women. Beyond the meaning conveyed in the words themselves, men tend to communicate with the objective of exchanging information and in doing so establish their social pecking order. This is why men tend to try to control the conversation by talking more than others and employing various methods to silence or demote those who disagree with them. Generally, men, unless aware of the need to do otherwise, also tend to center the conversation more on themselves (Anderson & Leaper, 1998; Coates, 1986). Again, this is a manifestation of using conversation as a way of establishing the social pecking order. Tannen (1994) classes this communicational behavior report talk. In contrast, women tend more than men do to be inclusive in their conversational styles. This is because women are more centered on creating rapport, rather than self promotion (Holmes, 1992; Johnson, 1993; Kilbourne & Weeks, 1997; Lakoff, 1975; Mulac, Erlandson, Farrar, & Hallett, 1998; Tannen, 1994; Tannen, 1995). Tannen (1994) classes this communicational behavior rapport talk.

Supporting this report versus rapport distinction, previous research has claimed a greater tendency by men, at least in oral conversations, to try to dominate (Herring, 1993; Holmes, 1992) and control the discussion (Edelsky, 1993), to be more competitive (Kilbourne & Weeks, 1997) and more assertive by interrupting others (Anderson & Leaper, 1998; West & Zimmerman, 1983; Zimmerman & West, 1975), and generally be more forceful (Weatherall, 1998) and less complementary (Coates, 1986; Yates, 2001). These gender-based differences in the cultural message imbued into the conversation are evident across cultures (Costa, Terracciano, & McCrae, 2001; Hofstede, 1980) and seem to also carry over to listserves (Herring, 1996b; Stewart, Shields, & Sen, 2001) and to e-mail in general (Boneva, Kraut, & Frohlich, 2001; Parks & Floyd, 1995). A direct consequence of these gender-based differences and preferences is that men and women tend to congregate into same-gender conversations (Tannen, 1994). Men talk more to other men than to women; women talk more to other women than to men. Interestingly, this happens also online in virtual communities where people have a much broader choice of communities to join and where they can hide their gender and identities or even masquerade as anything they wish to be known as (Gefen & Ridings, 2005).

The ITC in charge of discussions in online courses provides a unique opportunity to examine these cross-gender differences because it is possible to create what are arguably gender-neutral settings. Also, if, in these induced gender-neutral ITC settings, cross-gender differences exist, then these differences are probably not a matter of setting alone, but are a matter of the ingrained nature or nurture considerations extensively discussed in the cross-culture literature (Hofstede, 1980; Hofstede, et al., 1990) and should thus be controlled in the context of cross-cultural ITC research.

Whether cross-gender differences, such as gender congregation, apply also in online courses is actually an open question because parallels cannot be drawn with the closest equivalent, virtual communities, where these do apply (Gefen & Ridings, 2005). Virtual communities are not regulated by a moderator and people are free to come and go as they wish, without being graded on it. Moreover, in typical online courses, students do not interact with other students except in controlled threaded discussion settings where the teacher posts a question and the class then discusses it. This discussion is usually graded. The discussion is asynchronous, so it is impossible to dominate air time or control the discussion and who talks when as in oral discussions. Inserting socially loaded comments and body language cues is also impossible. It is as close as possible to a gender neutral setting. Moreover, in contrast to virtual communities where people join for many reasons, including the stereotypical feminine rapport and the stereotypical masculine information exchange, the reason people join online class discussions is
usually a matter of being forced to by the grading policy—a matter antitypical of both masculine and feminine stereotypical and sociolinguistic behavior. While gender-based communication patterns do occur in the regular classroom (Tannen, 1991), and some evidence does exist that men use the online environment more to access information and women more to converse (Herking, 1996a; Yates, 2001), how these apply to a supposedly gender-neutral setting, such as an online class, remains an open question. Should these gender-related communication behaviors carry over to these neutral settings, then it could be argued that they are another aspects of culture-induced behaviors.

HYPOTHESES

Accordingly, applying the underlying proposition that gender differences are not induced by the settings alone but rather are a matter of culture, then even in the relatively gender neutral ITC setting of online course discussions, some typical gender communication patterns should be evident. While the basic cross-gender difference of rapport versus report might be somewhat mute in these settings, other aspects, such as gender congregation, should still be evident. The rapport versus report distinction should be rather mute because online class discussions are deliberately not conductive for the feminine rapport type communication and are explicitly managed to discourage the male dominating status building communication styles. Moreover, the technical settings in these ITC, such as the asynchronous nature of online course discussions, do not permit the students to control who speaks, when, what they say, and for how long—again making aspects of stereotypical male alleged domination conversation styles immaterial (Tannen, 1995; West & Zimmerman, 1983).

And yet, other aspects of typical gender behavior should come through if the proposition holds. Primary among these is gender congregation during discussions. Men’s preference to respond to other men more than to women and vice versa could still come through even in these settings because there is nothing in the technical aspects of the ITC or in the way these conversations can typically be managed to exclude this possibility. There are no technical ITC aspects or plausible conduct rules that can make a student address comments, or not be able to address comments, by any other specific given student. Practically speaking, this should translate to a gender preference with men preferring to refer to other men and women to other women.

- **H1**: The number of references to postings by students of the same gender is higher than the number of references to posting of students from the opposite gender.

Although we do not expect students to resort strongly to their alleged stereotypical report versus rapport conversational styles in these ITC in general, some weaker aspects of these styles should still come through in conversational aspects, which are not forced by the ITC or typical course conduct regulations. One aspect in which these aspects should come through is in the extent of support given to positions presented by other students. Conventional political correctness in online courses may not be overly encouraging of blunt disrespect and challenging others, but there is a nuance students can play in whether they choose to be explicitly supportive of the postings of others or not. Extrapolating from the literature about typical gender conversational styles, and hence assuming these styles are culture induced and should therefore carry over also to gender neutral ITC settings: men should be less supportive of the positions of other students. Generally, men are supposed to be more assertive, competitive, and dominating (Anderson & Leaper, 1998; Edelsky, 1993; Herring, 1993; Holmes, 1992; Kilbourne & Weeks, 1997; Weatherall, 1998; West & Zimmer-
man, 1983; Zimmerman & West, 1975), and less complementary (Coates, 1986; Yates, 2001) than women. All these mount up to ways of shoring up one’s own social standing, a motive strong among men but rather absent among women (Tannen, 1994). This behavior should especially come through strongly when male students refer to other male students because, extrapolating from sociolinguistics, they should be competing with each other. When male students refer to postings by female students the competition should be one way, only by the male student.

• **H2:** Men referring to postings by other men will be less supportive than women are.

With female students, on the other hand, inclusion should be a more dominant feature of the conversation, as it is in other settings (Tannen, 1994). A central strategy in creating inclusion is showing support and encouragement toward the other person. If this carries over to gender neutral ITC, then it could be expected that women will be more supportive of other women because of their tendency to be inclusive among other women.

• **H3:** Women referring to postings by other women will be more supportive than men are.

**Method and Data**

The data for this study were extracted from online course discussions in 14 online courses. There was an average of eight online course discussions in each online course. Every one of these online discussions was analyzed. For each student in each online course discussion, we recorded how many postings there were, how many related to a previous posting by other students in this discussion, how many of these references to previous posting were to postings by male students, how many of these were supportive, how many were to postings by female students and how many of these were supportive. A posting was counted as supportive if the student posting it explicitly stated agreement or support with a previous posting in this discussion. We then removed those records that related to students who did not refer to postings by other students in this specific online discussion. These records were removed from the analyses because evidently these students were not taking an active part in the specific conversation but only posting to fulfill the course requirements. This left us with 599 records, dealing with 83 students who each participated on average in 7.2 online course discussions. Among these 599 records, 381 were of men who took an active part in the online course conversation and referred to postings by other students in the specific conversation and 218 were by women. The data were classified by two raters. On the overlapping sample of 100 posting, which was classified by both raters, there was absolute agreement.

Supporting the stereotype of men trying to control the conversation (Edelsky, 1993), men did significantly (T=2.751, p=.006) post more (mean = 3.00, standard deviation= 2.096) than women (mean = 2.55, standard deviation= 1.542) and did significantly (T=3.959, p<.001) post longer messages (mean number of words = 346.30, standard deviation= 205.903) than women (mean = 286.99, standard deviation= 154.767).

**Data Analysis**

To examine hypotheses H1 through H3, we compared the means of men and women with a set of T tests. Men in a given online course discussion did not significantly (T=1.067, p=.286) refer more to others (mean=.69, standard deviation=1.255) than women did (mean=.58, standard deviation=1.032). However, men did significantly (T=2.525, p=.012) refer more to other men (mean=.45, standard deviation=.913) than women did (mean=.28, standard deviation=.605), although women did not significantly (T=.692, p=.489) refer more to other women (mean=.28, standard deviation=.620) than
men did (mean=.24, standard deviation=.602). These results give partial support to H1. Gender congregation does occur, but primarily among men.

Surprisingly however, men in a given online course discussion were significantly (T=2.082, p=.038) more supportive of other men (mean=.14, standard deviation=.445) than women were (mean=.07, standard deviation=.254). This contradicts the expected direction in H2. Also, men did not significantly (T=1.544, p=.123) refer in a supportive manner to other students in general (mean=.20, standard deviation=.540) more than women did (mean=.13, standard deviation=.414). In fact, women were not significantly (T=.148, p=.882) more supportive of even only other women (mean=.06, standard deviation=.264) than men were (mean=.06, standard deviation=.289). This does not support H3. The hypothesized differences in the supportive behavior of students in online course discussions were not supported. Apparently, the courses were sufficiently gender neutral to make this otherwise typical behavior mostly insignificant.

We then examined if the gender of the student could be identified in a linear regression by the characteristics of postings the student made in the online course discussion. If this is so, it would lend more support to the claim that gender and communication style, also in these gender neutral settings, are related. In the linear regression, the gender of the student making the posting was the dependent variable. The length of the posting in words, whether this posting was supportive, and whether it was addressed to a student of the same or opposite gender, were the independent variables. In all, the explained variance was low at .02. The only significant determinant of student gender was the length in words of the postings the student made in this online discussion (β=.149, p<.001). However, when only the more active students were examined, the results became more convincing. When the analysis was limited to only those students who posted at least 3 postings in the conversation, there were 268 such records, the degree of explained variance became .15. The significant determinants were the length in words (β=.175, p=.004), the number of postings referring to previous postings by women (β=-.173, p=.006), and the number of postings referring to previous postings by men (β=.123, p=.048). In other words, among more students who participated more actively in the online course conversation, students who referred more to previous postings by men and less to previous postings by women were mostly significantly more likely to be men.

Discussion

Language is a central pillar of culture and subcultures within the dominant culture. It is a central pillar even within what may otherwise be considered the same national or historical culture. It is enough to read the famous words of George Bernard Shaw in *Pygmalion*, “An Englishman has only to open his mouth, in order to have another Englishman despise him,” to realize how even dialects create a cross-cultural event. This is a conclusion supported by research (Abrams & Hogg, 1987). Along those lines of brilliant eloquence, this study presents another aspect of cross-cultural communication, the *Vive la Différence*, according to which gender too is a central cultural difference.

As sociolinguistics claim, men and women apply language, and communication in general, to such a differing social objectives that cross-gender communications can be sometimes best seen as nothing less than cross-cultural miscommunications among people with differing cultural backgrounds (Tannen, 1994). Examining a derivative of this sociolinguistic viewpoint, this study hypothesized that even in the gender neutral ITC environment of online course discussions with their asynchronous and topic focused orientation, cross-gender communication would show some aspects of a cross-cultural communication. These hypotheses were partially supported, but,
the pattern in the data was strong enough to significantly allow the correct identification of the gender of the student participating in the online course discussion.

That gender should come through significantly in this otherwise deliberately gender and culture independent ITC setting lends support to the claim that there is a need to include gender as another significant aspect of culture, even within the same national culture environment. One should pay special heed to this conclusion because, to some extent, there is a voluntary gender segregation occurring in these discussions. This is something quite amazing when one stops to think about it, because it is occurring despite the gender neutral environment, which supports neither the typical male report type communication nor the typical female rapport type communication.

Before discussing these implications in detail, a word should be said about the limitations of the study. The data in this study came from a convenience sample. This is okay because the objective of the study was to show support to the need to include gender as an aspect of culture. Generalization was not the objective. Generalization requires replication in other and more varied ITC settings, including, but not limited to, other online courses and ITC supported business interactions. To this, one should add that no two courses are the same. Having said this though, the data of this exploratory study do warrant further investigation. Some gender behavior patterns did come through and did allow a significant identification of student gender.

So what do the data tell us? Gender, as sociolinguists and eminent playwrights tell us, is also about culture. Cross-gender communications have cross-culture aspects to them. While it is still unknown whether these gender differences are ingrained or learned, they did come through even when the settings, such as the one of this study, should have made them mute. When considering how culture affects ITC adoption and usage patterns, and it does (Gefen & Straub, 1997; Rose & Straub, 1998; Straub, 1994), this aspect should be considered too. Although more research is needed, sociolinguistics is one possible theory base to support this inclusion. On a practical level, these conclusions imply some interesting tentative implications. If men and women communicate with different objectives and so understand messages differently, then awareness and practical steps to address these misunderstandings should be taken both in online conversations and in other instances of ITC.

Explaining the gender effect, by focusing on the culture of language, may also explain some previous research results. People are generally more accepting of answers given by a computer generated cartoon when the topic of the answer provided through this cartoon corresponds to its gender stereotypes: male cartoons about sports and female cartoons about fashion (Lee, 2003). If gender preference is so much part of our everyday behavior that people show a tendency to congregate by gender even in gender neutral discussions, then this carryover of oral discussion gender behavior might explain why this happens. The results also provide additional explanations why women sense more social presence in business e-mail than men do (Gefen & Straub, 1997). Again the carryover of the respective gender aspects of communication to an ITC environment, which is supposed to be gender neutral, might explain this.

Looking at the results in a broader manner, the results of the study, if generalized, tentatively suggest that just as culture should be a major aspect of ITC research, so should gender. ITC research, and especially human computer interaction research, is about many things, but one of its central topics of research is about how people use ITC to communicate with other people, be it through e-mail, e-commerce, or virtual communities. A key aspect in such communication, determining its meaning and success, is the use of language. Since language cannot be understood properly when analyzed devoid of its social underpinnings, these socially overlaid meanings should
be part of any research on how ITC is used and how it supports communication among people. Ignoring these central social components, how they contribute to the meaning and value of ITC based communication, and especially how misunderstandings may arise when communicating across genders as they are across cultures, is tantamount to ignoring a central tenet of the ITC interaction process itself.

Including gender into human computer interaction research, however, requires a solid theoretical base. This lack of a strong theory base may explain why gender has not come up often before as a central aspect of this research. It is not enough to say there is a significant T statistic. There must also be a theory base, which can explain why there is this significance and so tie it into other research and a broader understanding. There might be undeniable physiological reasons which affect gender differences in ITC behavior, as some research suggests (Cutmore, Hine, Maberly, Langford, & Hawgood, 2000), but there are also cultural reasons, such as those presented in this study. These cultural psychological reasons are central in determining behavior. Sociolinguistics could be one theory base on which such understanding could be achieved. Looking at things through this theory could make our understanding richer and broader, and, more importantly, avoid a joint misclassification of men and women into one group, which ignores the different social meanings men and women attach to communication.

Unfortunately, while culture is recognized as a key issue in ITC adoption and usage patterns, most such research has chosen to ignore this aspect of gender. This may be because of political correctness constraints, but it is taking the unnecessary risk of being scientifically wrong. Smothering this cultural aspect not only hides significant relationships and blotches construct validity, but it also skews our understanding of the world. We know men and women have different managerial styles (Beasley, 2005; Boon, 2003) and handle domination and conflicts differently (Chan, Monroe, Ng, & Tan, 2006). We all know men and women think and communicate differently, whether saying so is or is not politically correct. It is about time ITC research also paid homage to gender. Ignoring gender may be a mistake, if gender differences come through even in the controlled settings of this study, how much more so that they should be evident in less controlled settings.

REFERENCES


Costa, P. T. J., Terracciano, A., & McCrae, R. R.


ENDNOTES

1 These differences are also related to genetics and hormones (Brizendine, 2006).

2 By the way, this is no joke. I tried it out on many of my students and almost unanimously all the men admitted to having been in this situation and driven around for hours, while almost all the women said they would ask for direction immediately.

Chapter 8.9
Growing Up Wireless:
Being a Parent and Being a Child in the Age of Mobile Communication

Letizia Caronia
University of Bologna, Italy

ABSTRACT

This chapter illustrates the role of the mobile phone in the rise of new cultural models of parenting. According to a phenomenological theoretical approach to culture and everyday life, the author argues that the relationship between technologies, culture, and society should be conceived as a mutual construction. As cultural artefacts, mobile communication technologies both are domesticated by people into their cultural ways of living and create new ones. How are mobile phones domesticated by already existing cultural models of parenting? How does the introduction of the mobile phone affect family life and inter-generational relationships? How does mobile contact contribute in the construction of new cultural models of “being a parent” and “being a child”? Analysing new social phenomena such as “hyper-parenting” and the “dialogic use” of mobile phones, the author argues upon the role of mobile communication technologies in articulating the paradoxical nature of the contemporary cultural model of family education.

BEYOND THE USER-TECHNOLOGY DICHOTOMY: A PHENOMENOLOGICAL APPROACH TO EVERYDAY LIFE

“Some day we will build up a world telephone system making necessary to all peoples the use a common language, or common understanding of languages, which will join all the people of the earth into one brotherhood” (Dilts, 1941, p. 11, cited in de Sola Pool, 1977, p. 129)

Like an underground current, the same social discourse reappears each time a new technology enters the social world: the technology purportedly produces new unexpected behaviours and causes major changes in the way people live. Whether it is for the worse or for the better is not
Growing Up Wireless

important. What matters more is the underlying unidirectional causal-deterministic model that putatively accounts for the influence of technologies in people’s lives.

The deterministic approach to social phenomena and particularly to technological evolution has had a long and strong tradition that spans the 20th Century. Even if today no one would say “science discovers, industry applies, man conforms,” the deterministic model persists in both scientific and commonsense approaches. At least within commonsense reasoning and theories, information and communication technologies are supposed to determine not only people’s behaviours but also their attitudes, relationships, and even identities. Empowered technologies are perceived as overwhelming unskilled people as if they dominate their lives. Such a view of the role of technologies in people’s everyday life has the hallmarks of all commonsense theories. It is self-evident, taken for granted, and ready made. It shares commonsense’s advantages: it provides easy to grasp explanations for a number of social events and allows people to cope with more dramatic circumstances. Like most practical reasoning, the one concerning information and communication technologies is a shortcut. It reduces the complexity of the phenomenon making it simpler and apparently more manageable.

Often echoed by media discourse and sometimes reinforced by references to simplified expert discourse, commonsense reasoning and layman theories constitute a shared cultural system through which we make sense of technologies in our daily life.

Although the deterministic approach to social phenomena has nurtured commonsense theories more than any other approach, it is not the only one. A major philosophical approach has been supporting concurrent views on social phenomena and providing a different paradigm for understanding technologies in everyday life: the phenomenological approach to social life.

Since Edmund Husserl’s and Alfred Schutz’s philosophical investigations, scholars in both Europe and the United States have emphasized the role of individuals in constructing culture, social organization, and their relation to the material features of everyday life contexts. Against any form of social and cultural determinism, ethnomethodology has demonstrated that people create their social and cultural world through their everyday actions and interactions (Garfinkel, 1967). Everyday practices of ordinary people are the effective tools that make supposedly passive users behave as active subjects. Defying and subverting any determinism of both dominant culture and the systems of production, social actors invent and create, moment by moment, the meaning and functions of things that circulate in their social space (De Certeau, 1984). Far from obeying implicit logics inscribed in goods, consumers develop their own tactics and follow paths in often unforeseen and unpredictable ways. The uses and gratification approach to information and communication technologies (Katz, Blumer, & Gurevitch, 1974) is consistent with this antideterministic paradigm. Proponents of this stream have shed light on the role of users’ needs and goals in the adoption or rejection of a technology and its intended uses.

These approaches to social life and phenomena share a crucial theoretical assertion: the strength of human agency (Giddens, 1979, 1984) and subject intentionality in making the meaningful dimensions of the world people inhabit. Accordingly, everyday life is conceived as a never-ending cultural work through which social actors produce the meaning, structures, and social organization of the world they live in, as well as their own identities and those of the people they interact with. Everyday language and interaction are the primary tools of this culture construction. However, social structures as well as the material features of everyday life contexts are more than an inert background for culture construction.
Disregarding any radical subjectivistic drift, the phenomenological approach to culture and everyday life does not underestimate the constraints of the world of things nor does it claim for an omnipotent actor. Rather it conceives the process of culture creation as radically embedded in the cultural frames and the material resources available in the world people inhabit, which in turn makes this creation possible.

As renewed attention to the material aspects of social life indicates (Appadurai, 1986; De Certeau, 1984; Gras, Jorges, & Scardigli, 1992; Latour, 1992; Semprini, 1999), the artifactual dimension of daily life is a crucial component that affects and is affected by everyday interactions, social organizations, and cultural frames of reference. Things, whether technological or not, participate in such a process of creating cultural models of living: as cultural artefacts, they are domesticated by users into pre-existing patterns of meaning and create new social scenarios and identities.

As people establish meaningful interactions with objects and artefacts, they make them exist in their social world, making sense of and domesticating them according to their frames of relevance and “moral economy” (Silverstone, Hirsch, & Morley, 1992). Literature on the social uses of media and the cultural ways of coping with a technological environment has shown how these uses, like other social practices, may be considered semiotic actions in the strict sense of the term; that is, ways of communicating and tools for constructing meanings and social realities.

The available technologies, the material features of the objects which support them and the daily routines they create or are integrated in, are all tools for the everyday production of culture and identities. Through media-related practices, individuals construct themselves in specific ways and produce the forms of their social participation (Caronia, 2002). Simply put, through our uses of media, through the way we act out these uses, we define (at least locally) the communities to which we belong and our identities.

We define but we are also defined. If face-to-face interaction and talk may still be considered the basic forms of socialization (Boden & Molotoch, 1994), the ways in which media uses become topics of everyday conversation are powerful tools to construct their meanings and the identities of those who use them. People’s ways of using media, whatever real or imagined, enter everyday conversations as parts of the narratives through which people constantly construct who they are and who the people they talk about are (Ochs & Capps, 1996).

If human beings construct the meaning of things and make sense of them according to their goals, the reverse is also true. Things are not neutral nor are they “pure” material objects waiting to be defined. Even though they do not determine people’s life, things delineate the conditions of possibility for new behaviours and ways of life. Their features and engineering anticipate paths of action and project new possible identities for the users. By moving the image from the permanence of the analogical universe to the ephemeral digital world, the digital camera demands and proposes a radical nonrealistic ontology for photography. Even the social perception of the photographer’s work and identity has changed. The digital camera has definitively legitimized photography as a manipulation of reality through iconic representation. Whereas the assumption of the nonreferential nature of documentary images has always been taken for granted by epistemologists and philosophers, the digital camera has integrated this representation of photography in the layman’s culture. Allowing people to make, remake and unmake iconic representations of reality, the digital camera has produced a new everyday culture of photography.

Overcoming the “subject-object” duality, we need to rethink the relationship between humans and technologies in terms of reflexivity, that is, a mutual construction of meaning and reciprocal sense making.
This need is even more pronounced for information and communication technologies. Their progressive introduction into people’s everyday life, the multiplication of possible new courses of action, and ways of communicating and getting information, expand the range of tools through which individuals construct culture and identities.

Faced with this changing and growing technological environment (Livingstone & Bovill, 2001), it then becomes relevant to investigate how the work of everyday culture construction may be affected by the new forms of technologically mediated actions, and vice versa.

The process of mutual construction among technologies, culture, and society may be analyzed at the macro level of patterns of diffusion and uses, as well as at the micro level of ordinary everyday interactions. Drawing upon data from qualitative and ethnographic research on mobile communication devices in ordinary life, this author of this chapter discusses the role of these technologies in the construction of family relationships and inner culture. Particularly, the chapter focuses on the following aspects: the creation of a cultural model of “parenting” (hyper-parenting), the dialogic use of mobile phones in connecting the different socio-cultural universes to which children belong (i.e., family and peer), and the role of mobile communication technologies in articulating the paradoxical nature of the contemporary cultural model of family education.

CONTEMPORARY STUDIES ON MOBILE PHONE DIFFUSION AND APPROPRIATION

In recent years, considerable research has examined the adoption and diffusion of mobile phone technology. It seems quite evident that, even though important differences exist across different countries (Kats & Aakhus, 2002), adolescents are a major well-established target for the adoption of this technology (Colombo, & Scifo, 2005; Kasesniemi & Rautianen, 2002; Ling, 1999; Lobet-Maris, 2003).

It is not surprising, then, that research has focused on young people’s uses of the mobile phone, especially in European countries where the adoption rate among adolescents and young people had been quite high. Investigating adolescents’ uses of mobile phone, Ling and Yttery (2002) show how adolescents hypercoordinate their social life and construct social encounters moment-by-moment. Mobiles phones allow for “perpetual contact” (Katz & Aakhus, 2002), a form of social link that seems to fit perfectly with young people’s peer culture and developmental tasks. Rather than voice calls, young people have made the Short Message System (SMS) their typical use of mobile phone (Cosenza, 2002; Grinter & Eldridge, 2001; Grinter & Palen, 2002; Riviére, 2002) Along with the economic advantages, the diffusion of the SMS among teenagers may be explained by social and cultural factors (Taylor & Harper, 2003). The silent dimension of this distant communication is at the core of its domestication in young people’s underground life both in the family and in school (Caron & Caronia, 2007). Allowing for silent and hidden communication, the mobile phone perfectly integrates a typical teenagers’ cultural pattern: constructing their social world outside of their parents’ control and the official rules governing life in school. Young people have also interpreted the technical constraints of SMS according to their specific peer culture. The limit of the numbers of available characters has been transformed into a resource for constructing a new language and new language games. Competence in this language defines the boundaries of a community of users, creates group membership and cohesion, and distances users from adults’ culture (idem).

Studying teenagers’ discourses on mobile phone, some scholars have noted that this technology is a detonator of social thinking: it provokes reflective thinking on the ethics, politeness, and
Growing Up Wireless

aesthetic rules of everyday action and social life (Caronia & Caron, 2004). Reflecting upon social uses of the mobile phone, teenagers explore the identity-making processes involved in the presentation of oneself on a public scene. They interpret and make the uses of the mobile phone work as a social grammar through which people supposedly define themselves and those around them. In this sense, using a mobile phone in a teenage-appropriate way is not a matter of technical competence; it requires broader communicative skills that include cultural knowledge of when, where, why, and especially how to use this technology. Similarly, researchers have analyzed the normative aspect of mobile phone use among teenagers’ groups. In particular, they focus on the implicit cultural rules governing the sharing of the technology (Caron & Caronia, 2007; Weilenman & Larsson, 2001). According to teenagers’ cultural frames of reference, the ownership of a communication device is not an individual matter. Rather it is a radically social affair. Alliance and friendship, leadership and membership, require sharing individual property: mobile phones are loaned and borrowed among the members of the group and this performance entails a system of reciprocal obligations. This “gift exchange” (Mauss, 1954 {1924}) reinforces social links and ritually defines who belongs to the group.

These studies shed light on different aspects of what can be conceived as a single process: the domestication and integration of the mobile phone into youth-specific culture. The mobile phone seems to work as a developmental tool that meets the needs of the growing up process. Particularly, young people use it to attain a certain degree of autonomy with respect to family world, to mark their belonging to a community of peers, to create their specific social organization, and to develop the skills and share the knowledge needed to become competent members of their own community.

Less explored than the world of teens are the cultural and social micro-aspects involved in parents’ uses of the mobile phone in communication with their children. Research on this issue mostly describes mobile phones’ usefulness in mutual coordination of children and working parents and their perception of mobile phone as a security/safety/control device (Caronia & Caron, 2004; Ling & Yttri, 2002; Rakow & Navaro, 1993). These studies have investigated relevant dimensions of the process through which mobile phones affect and are affected by family culture. However, more detailed knowledge and a deeper understanding of the cultural and interpersonal aspects of such a process is required.

Our hypothesis is that this mobile communication device contributes to the creation of new cultural models of being a parent and being a child.

CULTURAL MODELS OF PARENTING: A THEORETICAL APPROACH

The cognitive approach to culture conceives cultural models as prototypical, language-based scripts of events, actions, and social actors. These shared definitions of situations supposedly work as frames of reference for inference-making and as guides for appropriate, mutually understandable, and accountable actions. Shared by individuals belonging to the same linguistic and cultural community, cultural models are models of reality as they define what counts as an occurrence of what type of event. They are also models for reality insofar as they are used as references to act in accordance with these shared definitions of social events. Cultural models thus constitute a background cultural knowledge providing resources to understand and to perform in culturally appropriate manners (D’Andrade, & Strauss, 1992; Holland & Quinn, 1987). Just as linguistic competence and grammatical knowledge make linguistic performance and language use possible, the shared set of cultural models of reality gener-
mates social actions. Conversely, people’s actions, discourses, and behaviours are seen as merely reflecting or expressing their mental representations of the social world.

The cognitive approaches to social knowledge and praxis give primacy to knowledge over praxis, to culture over everyday actions and discourses. This top-down theoretical perspective has been strongly criticized and programmatically reversed by radical bottom-up views. Building upon philosophy of language and speech act theory (Austin, 1962; Searle, 1969), ethnomethodology (Garfinkel, 1967; Heritage 1984), social constructionism (Gergen, 1985), conversation analysis (Atkinson & Heritage, 1984), and discursive psychology (Edwards, 1997; Potter & Wetherell, 1987), conceive cultural models that organize everyday life in intersubjectively shared ways, as constructed moment by moment by the ways people participate in social events. Actions and discourses do not merely reflect an existing culture stored as information in people’s minds. Rather, they are tools for constructing cultural definitions of reality. From a radical constructivist perspective, knowledge is thus a product of praxis.

In contrast, dialectical perspectives have emphasized the reductionism of both the knowledge oriented and action oriented approaches. Phenomenology (Giorgi, 1990), cultural psychology (Bruner, 1996), critical discursive psychology (Wetherell, Taylor, & Yeats, 2001), and contemporary linguistic anthropology (Duranti, 1997) propose a theoretical perspective that captures the reflexive relationship between culture and action. Individuals are historical beings belonging to an existing life-world. This background of taken-for-granted assumptions, beliefs, and traditions provides established, normalized ways of understanding the world and sets the limits and the opportunities for acting and thinking (Foucault, 1980). However, background cultural knowledge does not determine people’s actions and behaviours. Through their everyday actions and discourses, individuals become crucial agents of a creative process of culture making, remaking, and unmaking. In essence, “knowledge and praxis create each other” (Ochs, 1988, p. 15).

Praxis is also technologically mediated actions and communications. How does the use of mobile communication devices shape cultural models of “being a parent” and “being a child” and vice versa?

**HYPER-PARENTING: A TECHNOLOGICALLY MEDIATED ACHIEVEMENT**

As the mobile phone became a tool for parent-child communication, it has been shaped by a pre-established culture of parenting. Pagers and mobile phones have been interpreted by parents as means to exert control over and fulfill their responsibilities toward their children. They have thus been completely domesticated in the family’s moral economy and transformed into tools for family socialization. By analyzing parents’ and children discourses on their use of mobile phones, we can reconstruct the repertoire of official reasons family members invoke to make sense of the adoption of this technology. Some patterns of meaning are recurrent, namely being in touch and responding to emergencies.

In the following example, Guy, a father in his mid-fifties, sees the mobile phone and pager as a kind of “umbilical cord,” since they allow him greater contact with his children:

**Guy:** But we also used it, now less, it used to be a lot like an umbilical cord with the kids. The kids could call us... Now it’s less important... they’re 19 and 20 now. They both have pagers. Bruno who didn’t want one, we twisted his arm to get him to have one, so we could get in touch with him.

Parents may also insist on their children calling them, as we can see in the following discussion between Louis (age 10) and his parents, Gerry
and Madeleine (in their 40s) that lend him their cell phone to reach them:

**Researcher:** Do you call your parents often?

**Louis:** Well, yes. Even when I’m going to school.

**Gerry:** Let’s say you don’t call, it’s because we tell you to call...

**Researcher:** Why do you call them, for example?

**Louis:** Well! Sometimes when it’s important or something that uh...

**Madeleine:** But he doesn’t call us. We have to insist on him calling us.

**Gerry:** In fact, it’s because we are starting to leave him at home alone a little. So we tell him, “before you go, you call.”

Children also can perceive the mobile communication device more as a kind of “electronic leash” that allows their parents to contact them at any time:

**Barry (19 years old):** It is a pager heu.. (...) and then afterwards there, it happened, what we call.. it becomes a bit like an electronic leash for my mother...

**Researcher:** An electronic leash?

**Barry:** For my mother and then so... It lets her call me all the time and then uh... any time.

**Andrée (mother, age 50):** Well, it’s true, I appreciate it.

**Barry:** Yeah, she finds it very useful.

**Andrée:** I can reach him everywhere because he has it and it’s reliable, you always have it?

**Barry:** I alw... I almost have it all the time on me. Sometimes I forget but otherwise it is always on, always, always.

The image of the cellular phone as a piece of emergency equipment is another recurrent pattern of meaning in parents’ accounts of the reasons they introduced cell phones. Parents and children often construct narratives of hypothetical dangers and imagined scenarios in which having a mobile phone helps the owner resolve a problematic situation. The emergency discourse is actually one of the most recurrent themes in explanations of why the mobile phone came into informants’ family and how they were supposed to use it:

**Researcher:** You told me you’d have a cellular?

**Louis (son, age 10):** Yes.

**Madeleine (mother age 40):** When did we talk to you about having a cellular?

**Louis:** When we were in the car.

**Gerry (father, age 45):** That’s right.

**Louis:** To call each other in emergencies.

**Gerry:** That is we will lend him one of our cell phones so that when he is on the mountain, if ever something happens, that he would get lost, like when he went to blue mountain and he got lost, well he would have his cellular and it is going to be programmed, because, you know, you can program the cell phone, so he will have to program the number.

**Louis:** At Green Mountain I had it.
Growing Up Wireless

Gerry: Had you lent it to him?

Madeleine: Yes I remember.

As these examples show, the use and the functions of mobile phones are shaped by typical features of cultural model of parenting characteristic of contemporary western society. Exerting control over children, ensuring that they are safe, handling emergencies, managing time to create family moments, assuming responsibility toward children, supervising children’s life out of the home, and mutual coordination to be in touch are all behaviours consistent with the cultural definition of being a parent. From this point of view, remote parenting (Caronia & Caron, 2004; Rakow & Navaro, 1993) seems to be nothing more than a new way to perform old functions and to act according to established models of fulfilling parental roles.

Our hypothesis is that mobile communication devices are not only an expression of an existing family culture and social organization, they are also ways to create them. By tracking their children’s movements, finding out who they are spending time with, claiming to know that their children are safe, reminding their children when they have to be back or scolding them if they are late, parents realize the rights and duties involved in “being a parent.” By participating in these remote parenting interactions, children are socialized in the commitments and responsibilities of being members of the family. Mobile communication devices’ practical uses are thus meaningful actions: they establish and confirm family boundaries, they state “who makes family with whom” and what behaviours belong to family members. They make the link between relatives permanent and work as teaching-learning strategies on the rights and duties governing family community life. Through the courses of action implicit in mobile phone use, parents do more than exert their role: they construct culture by legitimizing the definitions of what counts as “being a parent,” “being a child,” or “being a family” inscribed in their mobile phone mediated actions.

The following example from our ethnographic fieldwork on mobile use in the family sheds light on the cultural and social consequences hidden behind the most visible functions of remote parenting.

Scene: It is Saturday afternoon in Bologna. Silvia, a divorced mother in her 40s, is talking on the land phone to her friend, participant researcher Laura. Mafalda, Silvia’s oldest daughter, age 13, is in Milan at her father’s house. Silvia’s mobile phone rings:

1. Silvia to Laura: Wait, just a minute, it’s “Serafini-Milano.” (reading on the display)
2. Laura: Okay.
3. Silvia to Mafalda: Yes, sweetheart, I’m on the phone, talking with Laura.
5. Silvia: “Back?” (English in original), it means “at the rear” or “been returned,” It depends. You have to consider the sentence.
6. Silvia: “Appear?” (English in the original), I don’t know, wait, I’ll ask Laura...
7. Silvia to Laura on the land phone: Laura, what does it mean, “appear?” A-P-P-E-A-R?
8. Laura: “To have an outward aspect.”
9. Silvia to Mafalda: “To have an outward aspect.” Laura said that it means to have an outward aspect.
10. Silvia: “Fail” (English in original), I don’t know, I’ll ask Laura.
11. Silvia to Laura: And “fail?” (English in original), F-A-I-L?
12. Laura: To deceive, fail, to not succeed in doing something.
13. **Silvia to Mafalda:** To not succeed, okay sweetheart? Is it correct? I love you.

14. **Silvia to Laura:** Hey, many thanks, many thanks from Mafalda, too. It’s fantastic, I never do that, helping Mafalda with her homework in this way. She’s in Milano and I’m in Bologna and you, you’re at home, it’s great!

Remote parenting is more than an easy way to attain practical purposes or carry out typical functions related to the parental role. Through this mobile phone mediated interaction, Silvia is not only helping her daughter do her English homework. The sequence of her “mobile” actions is a meaning making devices.

Consider first the opening sequence of this multiparty telephone conversation. Silvia interrupts her conversation with Laura to give priority to her daughter’s call (turn 1.). “It is Serafini-Milano”: reading aloud the identity of the caller appeared on the display, she tells Laura whose needs come first. In this family culture “Serafini-Milano” is a shortcut for “daughters when they are at their father’s house.” Sharing this background local knowledge, Laura accepts Silvia’s shift to her daughter’s call (turn 2.). Acceptance is an action: through this action Laura legitimizes her being put in standby position. Then, Silvia does more than merely shift her attention to her daughter. After addressing Mafalda with some instances of intimate talk, she formulates this move with words (turn 3). Describing in words what one is doing is one of the linguistic moves through which participants negotiate the meaning of what is going on (Garfinkel & Sacks, 1970) and construct a shared definition of the event and its implications. This can be summarized as: “if children call, their mother is available and their needs come first.”

The second part of the conversation is a typical mother-child scaffolding interaction: the mobile phone guarantees a direct, always open access to the caregiver and makes it possible to carry out this kind of joint action despite physical distance.

In the closing sequence, Silvia formulates what happened as a new, original way to perform as a parent (turns 13 and 14).

The actions performed in and through this technologically mediated conversation are culture building activities and socialization devices. By doing “being a parent” in certain ways, the mother locally constructs and proposes dimensions defining her cultural model of parenting: being always available, giving priority to children’s needs over adults’ needs, interrupting the ongoing course of action to open up a parallel one, using intimate talk to address one’s own children, and giving children a scaffolding to overcome their difficulties.

At the same time, the mother’s actions convey a cultural model of the child: a child is a demanding individual whose needs come first. He or she has the right to expect his or her parent to divert attention from an ongoing adult-adult interaction to take care of the children.

What about the other participants? By participating in such an interaction, Mafalda is learning more than the meaning of some foreign words: she is being socialised in the cultural models of “being a mother” and “being a child” that are at stake in this family. Accepting her role as a collateral participant, Laura legitimizes the definition, under construction, of “what is going on and why.”

The participants officially involved in this interaction, are jointly constructing and ratifying, at least locally, a shared definition of “being a parent” and “being a child.” That is, they are constructing a cultural model of parenting as a technologically mediated achievement.

If “reciprocal availability” can be considered a trait of an existing culture of parenting, how does mobile phone shape this trait?

Mobile communication devices are distinctive because they allow anytime and anywhere reciprocal availability. This is nothing more than a suggested way of acting, a “possible world” inscribed in the technology. It is through everyday and ordinary ways of using the mobile phone that
this cultural model becomes a (technologically mediated) accomplishment.

A pre-existing cultural definition of being a parent has clearly shaped parents’ and children’s use of the mobile phone far beyond the management of urgency, safety, and control. In very reflexive ways, the use of this technology has created a new original way to be a parent. The possibility of remote parenting has been turned into hyper-parenting.

Through their everyday mobile interactions, parents and children have transformed physical distance into relational proximity; they have overcome the spatial and temporal constraints of face-to-face or traditional telephone interactions, and transformed almost every moment into an opportunity for coconstructing joint actions and care-giving. Family relationships seem to perfectly mirror the contemporary paradox of a wireless world producing hyperlinked people.

**LINKING MACRO AND MICRO: SOCIO-CULTURAL CHANGES IN FAMILY EDUCATION**

One of the most relevant contributions to the construction of the contemporary cultural model of parenting has been the historical change in the notion of parental authority. Since the end of the 17th Century, western societies have progressively moved from a political model of authority based on the concept of natural inequality and on the related notions of power and obedience to a definition of authority legitimised by a social consensus among equals. According to the political model, the pater familias’ supreme authority over his children was analogous to that exerted by a sovereign over his subjects. Power and submission were the expected behaviours defining parents’ and children’s reciprocal and naturally unequal status. This notion of authority was profoundly questioned by Locke and Rousseau according to the principle of the natural equality of all human beings. Obedience to the authority does not stem from the presumed inferior status of some individuals. Rather, it derives from the social contract. Individuals have to obey the law because it expresses the will of all the people concerned.

Since Toqueville, and throughout the 19th Century, this model of authority has strongly affected western societies’ macrohistory, politics, and laws. While in a slow often inconsistent way, it has also brought major changes to microhistory: it has affected the cultural models of all social relationships including those typical of the private family universe. The patria potestas no longer has a natural foundation, and negotiation is presumed to be at the core of family members’ distribution of authority. The translation of this consensus-based model into actions and interaction is less obvious with respect to the parent-child relationship. Here, a “natural inequality” (now referred to as “asymmetry”) is at stake: it defines the psycho-physical dependency of children on their caregivers. The gap in competences, skills, and knowledge is precisely what defines the reciprocal status of parents and children. Parents’ authority—whatever that means—is a tool to overcome this gap, a necessary condition for children’s development. Yet this necessity is no longer an argument for a power-based definition of parental authority.

At the end of the 20th Century, the inner logic defining parental authority was completely reversed. Until at least a century ago, parents had almost absolute power over their children, who were not considered to have rights. Now it is parents that have duties toward children. Most western societies have substituted the notion of patria potestas with those of parental cares and responsibilities, protection and supervision. The United Nations Convention on the Rights of the Child, introduced in 1989, emphasized that children have their own rights; adults’ “authority” is clearly defined in terms of duties, obligations, and responsibilities that parents have over their children. The notion of power as the probability
of obtaining obedience no longer defines the contemporary model of parenting.

How can this new image of a child as a subject that is no longer expected to obey be reconciled with the notion of parental authority and governance? How can authority be invoked without power? How does one cope with a child that is considered both a vulnerable and dependent individual and an autonomy-oriented person with rights?

Traditionally, parents had the right to exert power over children who had the obligation to obey. Aside from any other consideration, this model was inherently consistent. It offered a clear distribution of complementary rights and duties and traced a path for mutually coordinated actions.

The change in the definition of parental authority has broken the inner consistency of the traditional model. Now, children’s rights and parents’ duties often appear incompatible, and family education practices need to be conceived as ways to cope with the inherent paradoxes defining contemporary democratic education.

It may sound hazardous to relate these macro cultural, historical, and political changes to everyday life and social practices. Yet it is in the micro-order of ordinary life that cultural changes are both reflected and accomplished. The cultural changes in the status of children and in the notion of parental authority have directly and indirectly affected family education practices.

Aside from tragic exceptions, mainstream parents and children in contemporary western societies interact according to the social consensus framework. While recognizing the difficulties of creating such a consensus, sometimes shifting to or mythically evoking the simplicity of the authoritarian traditional model, contemporary parents have assumed the basic principles of democratic education.

Negotiation more than obedience, competence more than power, trust more than authority are the basic principles underlying parents’ and children’s reciprocal interactions. Mobile communication devices have been shaped by such a cultural framework. They have been appropriated as unexpected and efficient tools to negotiate parents’ and children’s often conflicting perspectives on children’s developmental needs, and as ways to articulate their often incompatible rights and obligations.

In other words, mobile communication devices play a role in coping with the slightly paradoxical nature of the contemporary model of family education.

PARENTS’ AND CHILDREN’S INCOMPATIBLE RIGHTS: MOBILE PHONES AS A NEGOTIATION TOOL

One of the domains where the paradoxical nature of parents’ and children’s reciprocal status emerges the most is children’s gradual shift towards the peer universe.

As they grow up, children multiply their universes of references and have to negotiate their progressive belonging to multiple socio-cultural worlds. Family and peer community are different, often conflicting, demanding worlds.

It is mostly at this point in their developmental trajectory that pre-adolescents start demanding autonomy, freedom, and the right to make their own decisions, choices, and even mistakes. Moreover, it is then that they construct social links beyond their parents’ filter. Not surprisingly, it is at this moment that parents officially enforce their responsibility, their obligation, or even their right to exert control and supervision. Less than 10 years ago, these conflicting perspectives were managed in two ways: talking with children about the where, when, what, and with whom of their extra-family life and negotiating the limits of time spent outside the home. These educational practices are rarely mutually exclusive. Depending on the family’s pedagogical model and on
children’s behaviours and attitudes, the transition to the peer universe may be more dialogic-oriented or more control-oriented. In either case, borders are clearly identified and the transition becomes a field where parents’ and children’s differing perspectives often conflict.

As mobile phones became part of the family’s technological equipment, they are interpreted as a means of coping with children’s belonging to multiple socio-cultural worlds. They have opened a new arena to manage the transition from family universe to that of peers. Crossing the boundaries of these two worlds, mobile phones become tools to negotiate children’s right to develop autonomy and parents’ right to control a still dependent child.

Consider first the arrival of a mobile phone in children’s life. Often, parents offer their children a mobile phone at some ritual milestone of their development. This gift-giving is a meaningful symbolic practice: parents recognize that their children have attained a stage of relative autonomy and allow them to privatize their social contacts. In contemporary societies, mobile phone gifts are rites of passage marking the (culturally perceived) beginning of adolescence. Like other goods provided by parents that symbolically mark the beginning of children’s autonomy, the mobile phone marks also children’s dependence on their parents for a wide set of needs and activities. Children’s dependence on their parents for mobile phone use seems even more obvious when we look at management of the related costs. As some adolescents, age 15, explain:

Researcher: Generally, among your friends who have mobile phones, who pays?

All: Our parents!

Mishan: Well, I think it is half and half because teenagers want to have a little independence so they pay half. But they can’t pay for the whole cost so their parents pay half.

Since children or even adolescents do not have the means to cover the costs, parents pay the fees. As soon as they are able, adolescents want to express their independence and give themselves some freedom by taking responsibility for part of the cost. Yet when parents are paying all or most of the related costs, they acquire the right to set rules on mobile phone use. This economic leash defines the boundaries of children’s autonomy. If the mobile phone is a bridge connecting children with the outside family universe, the bridge is built by parents who partially control access to this universe. Allowing for an economically controlled independency, mobile phones are used as cultural objects to mediate and gradually modulate children’s construction of a life outside of family boundaries and constraints.

Besides the strategic use of costs, the role of the mobile phone as a mediating device emerges from the way it is used both by parents and children.

As a 15-year-old girl told us, parents often take advantage of the fact that their children have mobile phones to keep an eye on them, watch over them from a distance, and even control where they go:

Karine (age 15): Well, it’s because my parents, it’s become like they really want to know where I am, so my parents always want to contact me, they’re always afraid when I’m outside. When I’m out they can’t contact me because there’s no phone, when I’m at someone’s it’s okay but if I go somewhere and I don’t call when I’m going, well they start freaking out, so… I need one.

Anxiety when children go out is an emotional experience strictly related to parenting. Mobile phones allow parents to cope with such a common experience in a totally original way: they do not need to choose between either allowing children be out of supervision and enduring their own anxiety, or radically limiting children’s social life to avoid painful apprehension. The mobile phone thus opens an alternative course of action:
**Pénélope (age 15):** Often there are, well I know some where it’s the parents who want to know where their kids are...

**Researcher:** Ah?

**Pénélope:** It’s a form of security for them, like, you know, parents buy it for their kids to give them, to give the kids some freedom, but at the same time the parents can know where the kids are any time.

For parents, the function at stake is not only to reduce their own apprehension. By tracking their children’s movements and finding out with whom they are spending time, they fulfil the duties and rights related to their parental role while respecting the children’s right to live an autonomous life. As we have seen in the extract above, adolescents seem to be quite aware of the dual nature of mobile phone contact. It is exactly this feature that allows them to use the technology in very strategic ways to negotiate their right to autonomy.

Adolescents usually accept that parents use cell phones to exert parental control. The duty to inform parents is generally well integrated in young people’s specific culture: this ritual practice is commonly assumed to be a way to gain more permission:

**Sophie (age 15):** It’s safer, not just for us, but for our parents too... Sometimes there are families that are separated; in ours, we’re all together, but even so my mother likes to be able to call me...

**Antoine (age 17):** In that sense, it’s sure that it’s another advantage, maybe to be able to get more permission to go out because you tell them something like “hey, you can call me any time, there won’t be any problems.”

**Sophie:** “If there’s anything, I’ll call you, Mom, Dad...”

**Karine (age 15):** Because some people go out a lot, they have a really big social life and for their parents, it would help them a little.

**Researcher:** To be able to contact them?

**Karine:** To contact them, yeah, so the parents won’t freak out, so they won’t be there going “Aaaack!”

As this discussion among adolescents shows, mobile phone has opened a totally new arena for negotiating children’s belonging to both family and peer worlds. Allowing users to reach and be reachable all time, the phone creates a symbolic space where parents and children can take into account the rights of the counterpart. When provided with a mobile phone, children are willing to accept the parents’ right to know where their children are, and they are willing to attend to and cope with their parents’ anxiety (It’s a form of security for them; It’s safer not just for us, but for our parents too.). Conversely, helping them manage their own right to apprehension, mobile phones help parents cope with their children’s right to go out and have a “really big social life:”

**Layla (age 15):** Why did you get a mobile phone?

**Jean (age 15):** Well, one, to communicate, like to have more independence. You know, when your parents tell you, like in a mall, you have something, a means of communicating, to contact you. Like with your friends, let’s say we say, “okay, let’s meet like at the movie theatre.”

**Layla (age 15):** Okay.

**Carl (age 15):** Me, it’s about the same too. It’s like the parent-child relationship. It’s sure that you know if, if you’re going to be late, you’re on your way, so then you just have to call your parents on
Growing Up Wireless

your mobile and then you tell them, “I’m coming, I’m about 15 minutes late, uh, I’m just in front of such-and-such a place, I’m coming.” Then they say, “oh, okay, that’s fine.”

Children consider autonomy, freedom, and peer life coordination the fundamental reasons for owning a mobile. However it is not surprising that they strategically focus on safety reasons when impressing on parents the need to get one. Grasping the unique opportunities they offer, adolescents use mobile phones as a means to act in ways that are consistent both with their needs and rights and with those of their parents.

Camera phones have amplified the repertoire of parents’ and children’s mediating tools. They provide parents and children with a more sophisticated strategy to negotiate their incompatible rights and to construct a bridge that connects the worlds of family and peers:

Layla (age 15): Then you say, “yes, Mom, see my friends,” then you take a picture with your phone. It’s so your mom can see that you’re really with your friends.

Researcher: Could you repeat what you just said?

Layla: It’s because I say to myself, okay, let’s say I call my mother, then my mother really wants to know if I’m somewhere or if I’m really with my friends, you take a picture, then you say, “yes, Mom, I’m with my friends, look.”

Researcher: So you’ve got proof.

Layla: Yeah, I’ve got proof.

Aside from their impressive role in creating a community of peer where instant pictures circulate and are shared by members, camera phones fit perfectly into parents’ and children’s typical interactions. It would be easy to invoke the image of Bentham’s panoptikon and wonder if this kind of visual remote parenting is nothing more than a contemporary version of power and control dynamics. Rather, we propose a different interpretation that focuses on the strategic use of this otherwise controlling device: the dialogic use of mobile phones. By letting their parents enter their peer world, by giving them visual proof of what they probably have stated in words, adolescents establish a contract of trust with their parents. Gaining in trust implies gaining in freedom and autonomy. Adolescents have turned the constraints of remote parenting to their advantages as in peer culture gaining more freedom is an advantage. Less hidden then it once used to be, adolescents’ underground life with respect to family can be strategically and at least partially shared with parents. The new sharable quality makes peer community experience a negotiable affair.

Sending photos via camera phone to parents may thus be seen as a paradox-resolving practice. In and through this process, parents and children articulate their symmetric and opposite rights, while creating much more room for negotiation and for consensus construction.

Mobile phones are also used to transgress this dialogic model. Transgressing family norms, pushing the boundaries of what has been consensually established as legitimate, and eventually breaking the terms of the trust contract, are all typical dimensions of adolescents’ growing up process. Mobile phones can be strategically turned off (i.e., when children do not want to be reached by their parents) or simply ignored if the caller ID device signals that parents are calling:

Tania (age 15): It’s like, I don’t really like that. You’re with your friends and then your mother calls you.

Researcher: So, you’d rather not have one? It’s your mother who insists?
Tania: Yeah, but it’s really like, uh, to know where I am, if I’m going to be late or something.

Researcher: So how do you deal with it?

Tania: I turn it off. [Laughs]

Researcher: You take it but you turn it off…

Tania: Well, I turn it off, yeah, or else I say I was in the subway.

The following is another example of the strategies used by adolescents to filter parental control calls. In this case an additional function of the technology enables Sandrine to transgress the rule:

Delphine (15 years old): On top of that, I don’t have call display, so I have no choice but to answer.

Sandrine (15 years old): I have call display, you know, when it’s someone, let’s say I’m somewhere and it’s my parents and I don’t want to talk to them, I don’t want them to bug me, so I don’t answer.

Turning off the mobile phone and not answering if parents are calling are behaviours that have to be justified or accounted for: no signal and low battery are the arguments commonly used by children to justify their being unreachable. The need for explanations defines these behaviours as exceptional transgressions. Invariably, transgression reveals and confirms the rule at stake. In this case, the rule is dialogic use of the mobile phone, a pattern of interaction legitimizing both parents’ right and duty to supervise their children and children’s right to be autonomous.

PARENTS’ AND CHILDREN’S DIALOGICAL USE OF MOBILE PHONES: CONCLUSIONS

Although they are incorporated in existing cultures of parenting and strongly dependent on each family moral economy, mobile communication devices nurture these worlds of meaning. Culture and praxis create each other. Everyday uses of mobile phone are no exception: they participate in the process whereby people constantly create and recreate their cultural ways of living their lives. Entering the family’s and teens’ everyday life and allowing new forms of interaction (such as remote parenting), this device works as a culture making object. It plays a major role in the rise of new cultural models of “being a parent” and “being a child.” Hyper-parenting is one such model.

Beside allowing coordination of children and working parents, besides facilitating single parents’ multitasking or connecting divorced parents to their children living away from their home, mobile phones are used to create and constantly confirm social and affective links among parents and children. Building upon traits of a shared, pre-existing model of parenting and seizing the opportunities of mobile phone, parents in contemporary western societies go beyond its practical functions. They have transformed the “reaching children and being reachable” dimension into anytime and everywhere reciprocal availability. Wireless parents and children perform as hyper-linked members of a family. Beyond any practical purposes or topic of mobile conversations, it is the contexts of mobile phone use and the ways in which it is used that give their interactions meaning. Their mobile verbal exchanges are meaningful actions that define their social and affective ties as relevant and prior to almost any other course of action. Simply put, mobile phones have magnified and overstated a totally cultural and unnatural parental model: children come first. It may sound obvious, yet it is not.
Growing Up Wireless

Mobile phones uses confirm, naturalize, and literally objectivize what is nothing more than a socially constructed definition of “being a parent.” These technologies participate in the silent and almost invisible process in and through which individuals create their cultural world as a quasi natural one.

The same process underlies the use of the mobile phone as a bridge connecting family and peer worlds. As we have seen, contemporary parents must contend with a definition of parental authority and children’s rights that make family education almost paradoxical. Mobile phone use has been totally integrated in a dialogical model of exerting parenthood. Thanks to its engineering, it leaves room for negotiating the symmetric and opposite rights defining the status of contemporary children and parents.

The mobile phone has opened a symbolic space to manage the transition from family universe to that of peers and to cope with paradoxes of the developmental process such as the typical autonomy-dependency dimension. Parents and children use the technology to blur the boundaries of the different socio-cultural universes to which children belong, and to make a smooth supported transition from family culture to the peer world. In particular, the mobile phone is used by family members to negotiate the often conflicting perspectives of parents and children with respect to the developmental needs and to articulate an often incompatible system of reciprocal rights and duties.

In a social and historical context where parental governance is no longer legitimized by power and obedience, mobile communication devices are used as an educational tool consistent with a consensus-based notion of parental authority. Performing according to such a cultural and even normative model of being a parent is a way to constantly recreate it. Through a dialogic use of mobile phones, parents and children participate in the process of defining a contemporary model of family education as a democratic social practice.

REFERENCES


Growing Up Wireless


ENDNOTES

1 This is the famous slogan of Chicago World's Fair in 1933.
2 The notion of intentionality as the human competence in making sense of reality and in creating the crucial dimensions of people’s life-world was introduced by Brentano and Husserl at the beginning of the 20th Century. It is perhaps one of the more heuristic notions of XX century philosophy.
3 Aside from any theoretical consideration, the relevance of talk-in-interaction in meaning making and culture construction has strong methodological consequences. Notably, it legitimizes the analysis of discourse as a social practice per se (Wetherell, et al., 2001). According to this framework, we conceive and analyze the uses of mobile communication devices as well as the discourses on such devices, as ways to make sense of them.
4 This chapter draws on data from a 7-year multiple research project on the appropriation of communication and information technologies in families' and adolescents' worlds (Caronia & Caronia, 2001, 2007; Caronia, 2005; Caronia & Caron, 2004). We used different yet complementary methodological and recording devices: family interviews and participant observation, log books on everyday practices, focus groups with adolescents, ethnographic case studies among natural groups of adolescents, analysis of naturally occurring mobile conversations and SMS exchanges. These studies have been conducted jointly with André H. Caron (Department of Communication, University of Montreal, Montreal), with financial support from CITÉ (Center for Interdisciplinary Research on Emerging Technologies, University of Montreal). I wish to thank my colleague André H. Caron for the support and the suggestions he gave me. Most of the ideas and interpretation proposed in this chapter have emerged from the joint analysis of data.
5 In some countries the diffusion and appropriation of mobile phone by young people is an important social phenomenon. In Europe, 23% of children ages 8 to 10 have a mobile phone. In Italy, one in three children between ages 5 and 12 has and uses a personal mobile phone. At age 9, 28% own a mobile phone. Among 14 and 18-year-olds, the percentage of mobile use and ownership is around 100% (Doxa Junior, 2005). In Norway, nearly 80% of children ages 9 to 12 have and use a mobile phone. The percentage rises at 13 (around 96%) to reach 100% at 16 (Ling, 2006). In Britain, 49% of children ages 8-11 and 82% of 12-15 year-olds own a mobile phone (Withers, 2006). The phenomenon is less apparent in Canada but is still growing. According to Statistics Canada, 17.5% of children in primary school have a cell phone and 42.8% of high school students have one (Dunfield, 2004).
6 The following verbatim come from family interviews made during a larger ethno-

2673
graphic research on the integration of new communication technologies in family everyday life (Caron & Caronia, 2001).

From an empirical point of view, we have strong evidences of remote parenting as a new way to perform the parental role. Whether it is considered as a way to dissimulate adults’ escape from their responsibility or as a way to manage the contemporary parents’ multitask life, strictly depends on the actors’ points of view and cultural models of “what counts as good parenting.” I advance the hypothesis that the common sense shared cultural model of “good parenting” is strictly anchored on notions as “face-to-face interaction” and “physical contact.” On the basis of such a cultural model, remote parenting may be considered as not consistent with the ideal-typical behaviour of a “good parent.” This is the point: are not these new communication technologies participating in the change of the cultural models we live by?

All the following transcripts come from conversational focus groups with adolescents on their uses of mobile phones (Caron & Caronia, 2004).

This work was previously published in Digital Literacy: Tools and Methodologies for Information Society, edited by P. Rivoltella, pp. 99-125, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
ABSTRACT

This chapter describes how social politeness is relevant to computer system design. As the Internet becomes more social, computers now mediate social interactions, act as social agents, and serve as information assistants. To succeed in these roles computers must learn a new skill—politeness. Yet selfish software is currently a widespread problem and politeness remains a software design “blind spot.” Using an informational definition of politeness, as the giving of social choice, suggests four aspects: 1. respect, 2. openness, 3. helpfulness, and 4. remembering. Examples are given to suggest how polite computing could make human-computer interactions more pleasant and increase software usage. In contrast, if software rudeness makes the Internet an unpleasant place to be, usage may minimize. For the Internet to recognize its social potential, software must be not only useful and usable, but also polite.

INTRODUCTION

Social Computing

Computers today are no longer just tools that respond passively to directions or input. Computers are just as mechanical as cars, but while a car inertly reflects its driver’s intentions, computers now ask questions, request information, suggest actions, and give advice. Perhaps this is why people often react to computers as they would to a person, even though they know it is not (Reeves & Nass, 1996). Miller notes that if I accidentally hit my thumb with a hammer, I blame myself not the hammer, yet people may blame an equally mechanical computer for errors they initiate (Miller, 2004). Software it seems, with its ability to make choices, has crossed the threshold from inert machine to interaction participant as the term human-computer interaction (HCI) implies. Nor are computers mediating a
social interaction, like e-mail, simply passive, as the software, like a facilitator, affects the social interaction possibilities (Lessig, 1999). As computers evolve, people increasingly find them active collaborators and participators rather than passive appliances or media. In these new social roles, as agent, assistant, or facilitator, software has a new requirement—to be polite.

To treat machines as people seems foolish, like talking to an empty car, but words seemingly addressed to cars on the road actually to their drivers. While the cars are indeed machines, their drivers are people. Likewise, while a computer is a machine, people “drive” the programs interacted with. Hence, people show significantly more relational behaviours when the other party in computer mediated communication is clearly human than when it is not (Sheetman & Horowitz, 2003), and studies find that people do not treat computers as people outside the mediation context (Goldstein, Alsio, & Werdenhoff, 2002)—just as people do not usually talk to empty cars. Reacting to a software installation program as if to a person is not unreasonable if the program has a social source. Social questions like: “Do I trust you?” and “What is your attitude to me?” now apply. If computers have achieved the status of semi-intelligent agents, it is natural for people to treat them socially, and thus expect politeness.

A social agent is taken as an interacting entity that represents another social entity in an interaction, either person or group, for example, if an installation program represents a company (a social entity), the installation program is a social agent, if it interacts with the customer on behalf of the company. The interaction is social even if the social agent is a computer, and an install creates a social contract even though the software is not a social entity itself. In the special case where a software agent is working for the party it is interacting with, it is a software assistant, working both for the user and to the user. In such cases of human-computer interaction (HCI), social concepts like politeness apply.

If software can be social it should be designed accordingly. A company would not let a socially ignorant person represent it to important clients. Yet, often, today’s software interrupts, overwrites, nags, changes, connects, downloads, and installs in ways that annoy and offend users (Cooper, 1999). Such behaviour is probably not illegal, but it is certainly impolite.

**Selfish Software**

The contrast to polite software is “selfish software.” Like a selfish person who acts as if only he or she exists, so selfish software acts as if it were the only application on your computer. It typically runs itself at every opportunity, loading at start-up and running continuously in the background. It feels free to interrupt you any time, to demand what it wants, or announce what it is doing, for example, after installing new modem software, it then loaded itself on every start-up and regularly interrupted me to go online to check for updates to itself. It never found any, even after many days, so finally after yet another pointless “Searching for upgrades” message I (first author) decided to uninstall it. As in “The Apprentice” TV show, one reaction to assistants that do not do what you want is: “You’re fired!”

Selfish software is why after 2-3 years Windows becomes “old.” With computer use, the Windows taskbar soon fills with icons, each an application that finds itself important enough to load at start-up and run continuously. Such applications always load, even if you never use them, for example, I never use Windows messenger but it always loads itself onto my taskbar. When many applications do this, it slows down the computer considerably, and taskbar icon growth is just the tip of the iceberg of what is happening to the entire computer. Because selfish programs put files wherever they like, uninstalled applications are not removed cleanly, and overtime Windows accretes an ever increasing “residue” of files and registry records left-over from previous installs. Giving
selfish applications too much freedom degrades performance until eventually only reinstalling the entire operating system can recover system performance.

**Polite Computing**

Polite computing is about how software design can support HCI politeness. It is not about how people should be polite to people online, which various “online etiquette” guides cover. This chapter aims to define, specify, and illustrate an information vision of polite computing.

Politeness is distinct from both usefulness and usability requirements. Usefulness addresses a system’s functionality, while usability concerns how people use that functionality. The first focuses on what the computer does, and the second on how the user gets the computer to do that. Polite computing, however, is not about what the computer does, nor how one can better get it to do it. It is about social relations rather than computer power or cognitive ease. It enables software that “plays well” in a social setting and encourages users to do the same. It addresses the requirements for social interaction, enabling better social collaboration, rather than better tool use. The contexts differ, so software could be easy to use yet rude, or polite but hard to use. While usability reduces training and documentation costs, only politeness lets a software agent work with a competent user without frustration. Both usability and politeness, however, fall under the rubric human-centred design.

**BACKGROUND**

The Oxford English Dictionary (http://dictionary.oed.com) defines politeness as:

... behaviour that is respectful or considerate to others

Considering and respecting others, a critical success factor in physical society, is equally relevant to online society. The predicted effect of polite computing is better human-computer interactions. While one may mistrust a polite door-to-door salesman as much as an impolite one, the polite one will get more “air time” because interacting with them is more pleasant. If politeness makes social interaction more pleasant, a polite society is a nicer place to be than an impolite one, and its people will be more willing to interact beneficially with others. Polite computing can contribute to computing by:

1. Increasing legitimate interactions.
2. Reducing anti-social attacks.
3. Increasing synergistic trade.
4. Increasing software use.

There is nothing to stop programmers faking politeness, just as nothing stops people in the physical world from doing so, but when people behave politely, cognitive dissonance theory finds they also tend to feel more polite (Festinger, 1957). Likewise, if programmers design for politeness, the overall effect will be positive, even though some may pretend.

**Politeness Supports Legitimate Interactions**

Legitimate interactions, defined as those that are both fair and in the common good, have been proposed as the complex social source of civilized prosperity (Whitworth & deMoor, 2003) and a core requirement for any prosperous and enduring community (Fukuyama, 1992). Conversely, societies where win-lose corruption and conflicts still reign are among the poorest in the world (Transparency-International, 2001). Legitimate interactions offer all parties a fair choice and are in the public good, while anti-social interactions, like theft or murder, give the “victim” little choice
Politeness as a Social Computing Requirement

Figure 1. The social choice dimension

<table>
<thead>
<tr>
<th>Polite</th>
<th>Legitimate</th>
<th>Anti-Social</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Degree of choice offered to the other party</td>
<td></td>
</tr>
</tbody>
</table>

and harm society overall. In contrast, polite acts are more than fair. To do as the law requires is not politeness precisely because it is required, for example, one does not thank a driver who stops at a red light, yet one thanks the driver who stops to let you into a line of traffic. While laws specify what citizens should do, politeness is about what they could do. If politeness involves offering more choices in an interaction than the law requires, then it begins where fixed laws end. If criminal acts fall below the law, then polite acts rise above it, and polite, legitimate, and anti-social acts can be ordered by the degree of choice offered to the other party or parties (Figure 1). In this view politeness increases social “health,” just as criminality poisons it.

Politeness Reduces Anti-Social Attacks

Polite computing may have value, but should not it take a back seat to security issues? Is politeness relevant if we are under attack? Yet upgrading security every time an attack exploits another loophole, is a never-ending cycle. An alternative is to develop strategies to reduce motivation to attack (Rose, Khoo, & Straub, 1999). Politeness can help one common source of attacks—resentment or anger against a system where the powerful are perceived to predate the weak (Power, 2000). Often hacking is vengeance against a person, a company, or the capitalist society in general (Forester & Morrison, 1994). Politeness contradicts the view that since everyone takes what they can, so can I. That some people are polite and give choice to others, may cause those neutral to society to copy, or those against society to become neutral. Politeness and security seem two sides of the same coin of social health. By analogy, a gardener defends his or her crops from weeds but does not wait for every weed to be killed before fertilizing. If politeness grows a better society, one should not wait to use it until every threat is purged. If security reduces anti-social acts, and politeness encourages social acts, they are complementary not mutually exclusive functions.

Politeness Increases Prosperity

Over thousands of years, as physical society became more “civilized,” this has created enormous prosperity, so for the first time in history some economies now produce more food than their people can eat (as their obesity epidemics testify). The bloody history of humanity seems to represent a social evolution from zero-sum (win-lose) interactions, such as war, to non-zero-sum (win-win) interactions, such as trade (Wright, 2001). Scientific research illustrates this social synergy, as for researchers to freely give their hard earned knowledge to all seems at first foolish, but when a critical mass do this, people gain more than they could have by working alone. Synergy means that when many people give to each other, they gain more than is possible by selfish activity. The success of the open source software (OSS) movement illustrates this, as open source products like Linux now compete with commercial products like Windows. The mathematics of social synergy are that while individual gains increase linearly with group size, synergy gains increase geometrically, as they depend on the number of interactions not the number of group members. The Internet illustrates social synergy, as we each only “sow” a
small part of it, but from it can “reap” the world’s knowledge interactions.

**Politeness Increases Software Use**

A study of reactions to a computerized Chinese word-guessing game found that when the software apologized after a wrong answer by saying “We are sorry that the clues were not helpful to you” the game was rated more enjoyable than when the computer simply said “This is not correct” (Tzeng, 2004). Brusque and often incomprehensible error messages like the “HTTP 404—File not Found” response to an unavailable Web page can imply a user fault, while a message like: “Sorry I could not find file xxxxx.” does not. Accusatory error messages can rub users the wrong way, especially if it is a software error in the first place.

In general, politeness improves the social interactions of a society, which makes it a nicer place to be. The reader can judge for him or herself whether the world wide Web is currently a nice place to be or whether its “dark side” which includes spam, spyware, viruses, hackers, pop-up ads, nagware, identity theft, solicitations, pornography, spoofers, and worms (Power, 2000), means it could benefit from polite computing. If software were more polite, people might be more willing to use it and less willing to abuse it.

**AN INFORMATION DEFINITION OF POLITENESS**

**Reinventing Politeness Online**

To apply politeness to computer programming, it must be defined in information terms. If politeness is “considering others,” then since different societies “consider” differently, what is polite in one culture can be rude in another. Given no universal “polite behaviour,” there seems no basis to apply politeness to the logic of programming. Yet while different countries have different laws, the goal of fairness that underlies the law can be attributed to every society (Rawls, 2001). Likewise, different cultures could have different “etiquettes” but a common goal of politeness. Figure 2 distinguishes the goals of Figure 1 from their specific implementations. In this view, while each society may “implement” a different etiquette, politeness remains the common “design goal,” just as legitimacy is the “spirit” behind laws that vary in detail between societies.

If politeness can take different forms in different societies, to ask which implementation applies online is to ask the wrong question. The right question is how to “reinvent” politeness in each specific online case, whether for chat, wiki, e-mail, or other groupware. Just as each different physical society develop local etiquettes and laws, so different applications may need a different politeness implementation based on a general design “pattern,” specifying politeness in information terms (Alexander, 1964).

**Informational Politeness**

If the person considered knows what is “considerate” for them, politeness can be defined abstractly as *the giving of choice to another in a social interaction*. Doing this is then always considerate if the other knows what is good for them, though the latter assumption may not always be true, for example, a young baby. In a conversation, where the locus of channel control passes back and forth between parties, it is polite to give control to the
other party, for example, it is impolite to interrupt someone, as that removes their choice to speak, and polite to let them finish talking, as they then choose when to stop.

An information definition of politeness is:

... any unrequired support for situating the locus of choice control of a social interaction with another party to it, given that control is desired, rightful and optional. (Whitworth, 2005)

Unrequired means the choice given is more than required by the law, as a required choice is not politeness. Optional means the polite party has the ability to choose, as politeness must be voluntary. Desired by the receiver means giving choice is only polite if the other wants it. “After you” is not polite when facing a difficult task. Politeness means giving desired choices, not forcing the locus of control, with its burden of action, upon others. Finally, rightful means that consideration of someone acting illegally is not polite, for example, to considerately hand a gun to a serial killer about to kill, is not polite.

**Other Definitions**

Some define politeness as “being nice” to the other party (Nass, 2004) and argue that when another says “I think I’m a good teacher; what do you think?” polite people respond “You’re great,” even if they do not think so. In this view, agreeing with another’s self praise is considered one of the “most fundamental rules of politeness” (Nass, 2004). Yet while agreeableness may often accompany politeness, it does not define it if one can be both agreeably impolite and politely disagreeable. One can politely refuse, beg to differ, respectfully object, and humbly criticize, that is, disagree but still be polite. Conversely, one can give charity to others yet be impolite, that is, be kind but rude.

Being polite is different from being kind, for example, kind parents may not give an infant many choices, but politeness does not apply to young children who are considered to not yet know what they really want. Do software creators consider software users to be like little children, unable yet to exercise choice properly? While inexperienced users may happily let software do as it thinks is best, when children grow up they want more choice (as teenagers illustrate). The view that “software knows best” is hard to justify for the majority of today’s computer-literate users. Perhaps once computer users were child-like, but today they want respect and choices from their software.

**Impolite Computing**

Impolite computing has a long history. Spam, for example, fills inboxes with messages users do not want (Whitworth & Whitworth, 2004) and is impolite because it takes choice away from e-mail receivers. Pop-up windows are impolite, as they “hijack” the user’s cursor or point of focus and take away the user choice of what they want to look at. Users do not like this, so many browsers prevent pop-ups. Impolite computer programs can:

1. **Use your computer’s services.** Software can use your hard drive to store information cookies or your long distance phone service for downloads.
2. **Change your computer settings.** Like browser home page, e-mail preferences or file associations.
3. **Spy on what you do online.** Spyware, stealthware, or software back doors that gather information from your computer without your knowledge or record your mouse clicks as you surf the Web and, even worse, exchange your private information with others.

For example, Microsoft’s Windows XP Media Player, was reported to quietly record the DVDs it played and use the user’s computer’s connection
to “phone home,” that is, send data back to Microsoft (Editor, 2002). Such problems differ from security threats, where hackers or viruses break in to damage information. This problem concerns those invited into our information home, not those who break in, for example, “software bundling,” where users choose to install one product but are forced to get many:

When we downloaded the beta version of Triton [AOL’s latest instant messenger software], we also got AOL Explorer—an Internet Explorer shell that opens full screen, to AOL’s AIM Today home page when you launch the IM client—as well as Plaxo Helper, an application that ties in with the Plaxo social-networking service. Triton also installed two programs that ran silently in the background even after we quit AIM and AOL Explorer. (Larkin, 2005)

Likewise, Yahoo’s “typical” installation of their IM also downloads their Search Toolbar, anti-spyware and anti-pop-up software, desktop and system tray shortcuts, as well as Yahoo Extras, which inserts Yahoo links on your browser. It also alters the users’ home page and auto-search functions to point to Yahoo by default. Even Yahoo employee, Jeremy Zawodny dislikes this:

I don’t know which company started using this tactic, but it is becoming the standard procedure for lots of software out there. And it sucks. Leave my settings, preferences and desktop alone. (http://jeremy.zawodny.com/blog/archives/005121.html)

A similar scheme is to use security updates to install new products, for example:

Microsoft used the January 2007 security update to induce users to try Internet Explorer 7.0 whether they wanted to or not. But after discovering they had been involuntarily upgraded to the new browser, they next found that application incompatibility effectively cut them off from the Internet. (Pallatto, 2007)

Security cannot defend against people one invites in, especially if it is the security system taking advantage! However, in a connected and free society, social influence can be very powerful. In physical society the withering looks given to the impolite are not toothless, as what others think of you affects how they behave towards you. In old societies banishment was often considered worse than a death sentence. Likewise, what online users think of a company that creates a software agent can directly impact sales. A reputation for riding roughshod over computer user’s rights is not good for business.

**SPECIFYING SOFTWARE AGENT POLITENESS**

The widespread problem of software that is rude, inconsiderate, or selfish is a general software design “blind spot” (Cooper, 1999). The specification of politeness in information terms is in its infancy, but previous work (Whitworth, 2005) suggests polite software should:

1. **Respect the other’s rights.** Polite software respects the user, does not pre-empt user choices, and does not act on or copy information without its owner’s permission.
2. **Openly declare itself.** Polite software does not sneak or change things in secret, but openly declares what it does, who it represents, and how they can be contacted.
3. **Help the other party.** Polite software helps users make informed choices, giving useful and understandable information when needed.
4. **Remember the interaction.** Polite software remembers past user choices in future interactions.
Respect includes not taking another’s rightful choices. If two parties jointly share a resource, one party’s choices can deny the other’s, for example, if I delete a shared file, you can no longer print it. Polite software should not preempt rightful user information choices regarding common HCI resources like the desktop, registry, hard drive, task bar, file associations, quick launch, and other user configurable settings. Pre-emptive acts, like changing a browser home page without asking, act unilaterally on a mutual resource and so are impolite.

Information choice cases are rarely simple, for example, a purchaser can use the software but not edit, copy, or distribute it. Such rights can be specified as privileges, in terms of specified information actors, methods, and objects (Table 1). To apply politeness in such cases requires a legitimacy baseline, for example, a software provider has no right to unilaterally upgrade a computer the user owns (though the Microsoft Windows Vista End User License Agreement (EULA) seems to imply this). Likewise, users have no right to unilaterally upgrade, as this edits the product source code. In such cases politeness applies, for example, the software suggests an update and the user agrees, or the user requests an update and the software agrees (for the provider). Similarly, while a company that creates a browser owns it, the same logic means users own data they create with the browser, for example, a cookie. Hence, software cookies require user permission, and users should be able to view, edit, or delete “their” cookies.

A respectful assistant does not interrupt unnecessarily, while selfish software, like a spoilt child, repeatedly does, for example, Windows Update advises me when it starts, as it progresses, and when it finishes its update. Its modal window interrupts what I am doing, seizing the cursor, and loses my current typing. Since each time Update only needs me to press OK, this is like being repeatedly interrupted to pat a small child on the head. The lesson of Mr. Clippy, that software serves the user not the other way around, seems still unlearned at Microsoft.

It is hard for selfish software to keep appropriately quiet, for example, Word can generate a table of contents from a document’s headings. However, if one sends the first chapter of a book to someone, with the book’s table of contents (to show its scope), every table of contents heading line without a page number loudly declares: “ERROR! BOOKMARK NOT DEFINED.” This, of

<table>
<thead>
<tr>
<th>Actors</th>
<th>Objects</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>Persona (represent people)</td>
<td>Create/Delete/Undelete</td>
</tr>
<tr>
<td>Groups</td>
<td>Containers (contain objects)</td>
<td>Edit/Revert</td>
</tr>
<tr>
<td>Agents</td>
<td>Items (convey meaning)</td>
<td>Archive/Un-archive</td>
</tr>
<tr>
<td></td>
<td>- Comments (dependent meaning)</td>
<td>View/Hide</td>
</tr>
<tr>
<td></td>
<td>- Mail (transmit meaning)</td>
<td>Move/Undo</td>
</tr>
<tr>
<td></td>
<td>- Votes (choice meaning)</td>
<td>Display/Reject</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Join/Resign</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Include/Exclude</td>
</tr>
</tbody>
</table>
course, completely spoils the sample document impression, and even worse, this is not apparent until the document is received. Why could the software not just quietly put a blank instead of a page number? Why must it announce its needs so rudely? What counts is not what the software needs, but what the user needs.

Open

Part of a polite greeting in most cultures is to introduce oneself and state one’s business. Holding out an open hand, to shake hands, shows that the hand has no weapon and that nothing is hidden. Conversely, to act secretly behind another’s back, to sneak or to hide ones actions, for any reason, is impolite. Secrecy in an interaction is impolite because the other has no choice regarding things they do not know about. Hiding your identity reduces my choices, as hidden parties are un-touchable and unaccountable for their actions. When polite people interact, they declare who they are and what they are doing.

If polite people do this, polite software should do the same. Users should see who is doing what on their computer. However, when Windows Task Manager shows cryptic process like CTSysVol.exe, attributed to the user, it could be system critical process or one left over from a long un-installed product.

An operating system Source Registry could link all online technical processes to their social sources, giving contact and other details. “Source” could be a property of every desktop icon, context menu item, taskbar icon, hard drive file, or any other resource. A user could delete all resources allocated by a given source without concern that they were system critical. Windows messages could also state their source so users know who a message is from. Source data could be optional, making it backward compatible. Applications need not disclose themselves, but users will prefer sources that do. Letting users know the actions of their computer’s inhabitants could help the marketplace create more polite software.

Helpful

A third politeness property is to help the user by offering understandable choices, as a user cannot properly choose from options they do not understand. Offering options that confuse is inconsiderate and impolite, for example, a course text Web site offers the choices:

- OneKey Course Compass
- Content Tour
- Companion Web site
- Help Downloading
- Instructor Resource Centre

It is unclear how the “Course Compass” differs from the “Companion Web site,” and why both seem to exclude “Instructor Resources” and “Help Downloading.” Clicking on these choices, as is typical for such sites, leads only to further confusing menu choices. The impolite assumption is that users enjoy clicking links to see where they go. Yet information overload is a serious problem for Web users, who have no time for hyperlink merry-go-rounds.

Yet to not offer choices at all, on the grounds that users cannot understand them, is also impolite. Installing software can be complex, but so is installing satellite TV technology. In both cases users expect to hear their choices in an understandable way. Complex installations are simplified by choice dependency analysis of how choices are linked, as Linux’s installer does. Letting a user choose to install an application they want minus a critical system component is not a choice but a trap. Application-critical components are part of the higher choice to install or not, for example, a user’s permission to install may imply access to hard drive, registry, and start menu, but not to desktop, system tray, favourites, or file associations.
Remember

Finally, it is not enough to give choices now but forget them later. If previous responses are forgotten, the user must redo them, which is inconsiderate. Hence, software that actually listens and remembers past user choices is a wonderful thing. Polite people remember previous encounters, yet each time I open Explorer it fills its preferred directory with files I do not want to see, then returns the cursor to me to select the directory I want to look at, which is never the one displayed. Each time, Explorer acts as if it were the first time I had used it, yet I am the only person it has ever known. Why can it not remember where I was last time and return me there? The answer is simply that it is impolite by design.

Such “amnesia” is a trademark of impolite software. Any document processing software could automatically open the user’s last document and put the cursor where they left off, or at least give that option (Raskin, 2000). The user logic is simple: “If I close the file I am finished, but if not, put me back where I was last time.” Yet most software cannot even remember what we were doing last time we met. Even within an application, like Outlook’s e-mail, if one moves from inbox to outbox and back, it “forgets” the original inbox message and one must scroll back to it.

If a choice repeats, to ask the same question over and over, for the same reply, is to pester or nag like the “Are we there yet?” of children on a car trip. This forces the other party to again and again give the same choice reply, for example, uploading a batch of files creates a series of overwrite questions, and software that continually asks “Overwrite Y/N?” forces the user to continuously reply “Yes.” Hence, most copy software also offers the “Yes to All” meta-choice that remembers for the choice set. Offering choices about choices (meta-choices) reduces information overload, as users need only set repeated access permissions once, for example:

1. Always accept
2. Always reject
3. Let me choose

A general meta-choice console (GMCC) would give users a common place to see or set all meta-choices (Whitworth, 2005).

IMPLEMENTATION CASES

The Impolite Effect

In HCI interactions, impoliteness can cause a social failure every bit as damaging as a logic failure, for example, the first author’s new 2006 computer came with McAfee Spamkiller, which when activated overwrote my Outlook Express mail server account name and password with its own values. When checking why I could no longer receive mail, I retyped in my mail server account details and fixed the problem. However, next time the system rebooted, McAfee rewrote over my mail account details again. The McAfee help person explained that Spamkiller was protecting me by taking control and routing all my e-mail through itself. To get my mail I had to go into McAfee and tell it my specific e-mail account details. That this did not work is less the issue than why this well known software:

a. Felt entitled to overwrite the e-mail account details a user had typed in
b. Could not copy my account details, which it wrote over, to create its own account

deep this same software also “took charge” whenever Outlook started, forcing me to wait as it did a slow foreground check for e-mail spam. Yet in 2 weeks of use, it never found any spam at all! I (first author) concluded it was selfish software, and uninstalled it.
Politeness as a Social Computing Requirement

Interaction Situations

Other human computer interactions where politeness applies include:

1. **Errors.** Polite error messages say we have an error rather than you have an error. While computers tend to take charge when things go well, when they go wrong, software seems to universally agree that the user is in fact “in charge.” To ask what “we” (rather than you) want to do about an error implies the computer should also suggest solution options. Studies of users in human-computer tutorials show significant differences based on how politely the computer addresses the user, that is, users respond differently to “Click the Enter button” vs. “Let’s click the Enter button” (Mayer, Johnson, Shaw, & Sandhu, 2006).

2. **Advice and Notifications.** To interrupt impolitely disturbs the user’s train of thought. For complex work, like programming, even short interruptions can cause a mental “core dump,” as the user drops one thing to attend to another. The real interruption effect is then not just the interruption time, but also the user recovery time (Jenkins, 2006), for example, if a user takes three minutes to refocus after an interruption, a 1 second interruption every 3 minutes can reduce productivity to zero. Mr. Clippy, Office ‘97’s paper clip assistant, had this problem, since as one user noted: “It wouldn’t go away when you wanted it to. It interrupted rudely and broke your train of thought.” (Pratley, 2004). Searching the Internet for “Mr. Clippy” gives comments like “Die, Clippy, Die!” (Gauze, 2003), yet its Microsoft designer wonders: “If you think the Assistant idea was bad, why exactly?” (Pratley, 2004). To answer simply, he was impolite, and in XP, is replaced by polite smart tags.

3. **Action requests.** Asking permission is polite because it gives the other choice and does not pre-emptively act on a common resource, such as a zip extract product that puts the files it extracted as icons onto the desktop, without asking! Such software tends to be used only once.

4. **Information requests.** If software asks for and gets choices from a user, it should remember them. Polite people do not ask “What is your name?” every time they meet, yet software often has no interaction memory whatsoever, for example, when reviewing e-mail offline in Windows XP, actions like using Explorer trigger a “Do you want to connect?” request every few minutes. No matter how often one says “No!” it keeps asking, because the software has no interaction memory.

5. **Installations.** Installation programs are notorious for pre-emptive acts, for example, the Real-One Player adds desktop icons and browser links, installs itself in the system tray, and can commandeer all video and sound file associations. Customers resent such invasions, which while not illegal, are impolite. An installation program changing your PC settings is like furniture deliverers rearranging your house because they happen to be in it. Software upgrades continue the tradition, for example, Internet Explorer upgrades that make MSN your browser home page without asking. Polite software does not do this.

Online Learning

Online learning software, like WebCT or Blackboard, illustrates how politeness issues vary with channel type. While channel richness (rich vs. lean) was once thought the main property of computer-mediated communication (Daft & Lengel, 1986), channel properties like linkage (one-to-one,
Politeness as a Social Computing Requirement

One-to-few or one-to-many) and interactivity (one-way or two-way) now also seem relevant (Whitworth, Gallupe, & McQueen, 2001). For example, instructor-student online communications, like e-mail, text messaging, chat, podcasts, cell phone, or video-computer interaction are usually one-to-one and two-way. In contrast, instructor-class communications are one-to-many and one-way. The rich-lean dimension is orthogonal to this distinction, for example, an instructor can post lean text assignments, graphical lecture slides, or rich video-lessons. E-mail still plays a major role in online learning, though it remains largely plain text, because it is interactive. Online learning system’s e-mail and chat functions unnecessarily duplicate existing e-mail services, like Hotmail. Having a separate e-mail for each class taken or taught requires students or instructors to check each class e-mail, in addition to their normal e-mail. For online learning systems to create normal e-mail lists would be much more user considerate, as then students would only have to check their normal e-mail.

In 1:1 two-way communications, like e-mail, “the conversation channel” is the shared resource. Yet while physical society recognizes the joint ownership of communication channels and offers everyone the right not to interact (e.g., to remain silent, to not receive junk mail, to not answer the phone, etc.), the core e-mail system gives all senders the right to put any message into any receiver’s inbox. This unfairly gives all rights to the sender and none to the receiver and enables the ongoing spam epidemic that plagues us all.

A more fundamental problem with e-mail in online learning is that one-to-one teacher-student interactions do not scale well (Berners-Lee, 2000). While one can as easily post lessons to a large class as to a small one, handling e-mails for classes over 50 can be difficult. The legitimacy baseline is that students have paid for class tuition, not one-to-one on-demand tuition. Experienced instructors often restrict the use of e-mail to personal requirements, like arranging meetings. They discourage its use for course content, for example, “Sorry I could not make the last class, what did I miss?” is a real student e-mail that I discouraged. Politeness in an interaction works two-ways, so training students to be e-mail polite is a valid learning goal, for example, polite e-mails are:

1. **Signed.** Give your name clearly—e-mails from nicknames like “fly-with-wind” are often unanswered.
2. **Understandable.** Give course/class number in the e-mail title so the instructor knows the context.
3. **Personal.** Use personal e-mail for personal issues, not issues that affect the entire class, for example, an online instructor may paste a “When is the exam?” e-mail into an online discussion board and answer it there, so other students can see the answer.

Class to instructor interactions, like an online assignment submission box, illustrate many-to-one one-way communication. For multi-choice quizzes, the computer can also grade the submissions and give student feedback. This is scalable as the computer can handle any class size, and can remember previous tests, telling the student if he/she is improving or not. However, while online exams do not need politeness, as students must take them, voluntarily online learning is a different matter. The distinction is:

a. **Formal testing quizzes.** Usually begin and end at a fixed time, shuffle questions and options to prevent cheating, and give little content feedback. Being mandatory, politeness applies only minimally.

b. **Informal learning quizzes.** Offer choices like pausing to restart later, optional tips, answer feedback, and choice of difficulty level. Being voluntary, politeness can help involve the student in the learning process.
Politeness as a Social Computing Requirement

If learning means changing one’s own processing, a case can be made that all learning is voluntary. If so, polite interaction may help engage students in voluntarily online learning. The difference between a forced online quiz and an online learning experience may be politeness and respect. Online quizzes can support face-to-face lessons, for example, if students answer online questions on a textbook chapter the week before lecture. This questioning encourages them to actively find information from the textbook, and prepares them for the weekly face-to-face class. Unlike a testing quiz, which is given after the class, and is graded by percentage correct, a “learning participation” quiz occurs before the taught class, and any reasonable participation (e.g., 30+) gets full points. However, the quiz must be done in the week stated, and there are no “resits” for weekly participations. The quiz answers are not released until the week finishes, and students can do or redo the quiz any time in the given week. In practice, those who do poorly in testing quizzes also tend to omit the learning quizzes. However, the good students find them an excellent way to learn.

Most online learning systems seem designed to give information to teachers rather than students, who get learning feedback only with difficulty, for example, Figure 3a shows a “View Scores” button, which when clicked gives Figure 3b, that shows a score. Few students then realize that clicking the underlined “1” gives feedback on the right answers. While online teachers can “see” everything, like when and for how long students are online, students struggle to see what could help them learn in online software.

Figure 3. Getting quiz feedback in WebCT
Class-to-class FAQ boards, where students answer each other’s questions, are many-to-many, two-way interactions that scale well to all class sizes. Respecting and using class member knowledge is not only popular with students but for fast changing subjects, like Web-programming, almost essential. If young people learn mainly from their peers, involving their peers in online learning seems sensible, and polite computing could enable this.

Polite computing suggests voluntary choice is a new online learning dimension. Its application however, requires a complete redesign of current teacher focused systems like WebCT. The online classroom must move from what is essentially a software supported dictatorship to a system that invites voluntary student participation, based on a balance of rights and choices.

FUTURE TRENDS

Polite computing suggests computers will increasingly:

1. Remember interaction data rather than object data.
2. Become human assistants or agents rather than independent actors.
3. Support politeness rather than selfishness in online interaction.

Remember the Interaction

It is astounding that major software manufacturers like Microsoft gather endless data on users, but seem oblivious to data on how their software interacts with the user. Like Peter Sellers in the film “Being There,” such software “likes to watch,” but cannot relate to people. To spy on users at every opportunity is not a user relationship. For example, Mr. Clippy watched your document actions but could not see his interactions with you, and so was oblivious to the rejection and scorn he evoked. Most software today is in the same category, and modern airport toilets seem more aware of their users than the average personal computer. Hopefully tomorrow’s software will make HCI memory its business, as its primary role will be to work for people, not for itself.

Computers as Assistants or Agents

There are several reasons why people should control computers, not the reverse. Firstly, while computers manage vast amounts of data with ease, they handle context changes poorly, and outside their fixed parameters can seem very stupid. So-called “smart” computing (Kurzweil, 1999) usually needs a human “minder.” Secondly, computers are not accountable for what they do, as they have no “self” to bear any loss. If society makes people accountable for what computers do, as it does, people need control over computer choices. Thirdly, the resistance of people to computer domination is predictable. Software designers should not underestimate the importance of user choice. In human history, freedom and choice are the stuff of revolutions, and a grass-roots Internet movement against impolite software is not inconceivable.

Fortunately, the future of computers probably lies not in becoming so clever or powerful that people are obsolete, nor in being passive human tools, but in contributing to a human-computer combination that performs better than either people or computers alone. The runaway IT successes of the last decade (cell-phones, Internet, e-mail, chat, bulletin boards, etc.) all support people rather than supplant them. As computers develop this co-participant role, politeness will be a critical success factor. These arguments suggest that if the role of computers is to assist, they should learn to be polite.
Online Politeness Will Grow

Today, many users feel at war with their software: removing things they did not want added, resetting changes they did not want changed, closing windows they did not want opened, and blocking e-mails they did not want to receive, and so forth. User weapons in this unnecessary war include third party blockers, cleaners, and filters of various sorts, whose main aim is to put users back in charge of their computer estate. Such applications are the most frequent accesses at Internet download sites. Like all wars, if software declares war on user choice, everyone will lose in the long run. If the Internet is a battlefield, no-one will want to go there. Some compare the Internet to the U.S. Wild West, and others talk of the “hunter-gatherers of the information age” (Meyrowitz, 1985). Yet the Stone Age and the U.S. Wild West evolved into civil society, and so perhaps it is time to introduce civility to the Internet. What took physical society thousands of years may occur online in only a few years for example, Wikipedia began with few rules and one leader, but now to combat “trolls” who trash data, has many rules (including copyright) and many roles, like “Steward,” “Bureaucrat,” and “Sysop” (Whitworth, Aldo de Moor, & Liu, 2006). Yet the real force behind Wikipedia is the majority’s enjoyment of working together considerately, not its ability to deal with the anti-social minority.

Many successful online traders find politeness profitable. eBay’s customer reputation feedback gives users optional access to valued information relevant to their purchase choice, which by the previous definition is polite. Amazon gives customers information on the books similar buyers buy, not by pop-up ads but as a view option below. Rather than a demand to buy, it is a polite reminder of same-time purchases that could save the customer postage. Politeness is not about selling but improving the customer relationship that leads to sales. By giving customers choice, polite companies win business because customers given choices come back. Perhaps one reason the Google search engine swept all before it was that its simple white interface, without annoying flashing or pop-up ads, made it pleasant to interact with. Google ads sit quietly at screen right, as options not demands. Yet while many online companies know that politeness pays, for others the lesson is still being learned, and for still others, hit-and-run rudeness is an online way of life.

FUTURE RESEARCH

The users of modern software increasingly choose whether to use it or not, for example, President Bush’s 2001 decision not to use e-mail because he did not trust its privacy. The ability of software to hold users hostage to its power may be declining. Where customers choose their software, a simple prediction is made: Polite software will be used more and deleted or disabled less than impolite software.

An experimental test of polite computing value requires a comparison of polite versus impolite applications on measures like willingness to use, attitude to the software, willingness to purchase, and user satisfaction. Politeness here is defined to apply not just to language, conversations, or people, but also to human-computer interactions. Research can show if computer users really value politeness in HCI interactions like application installations, user help, online learning, e-mail, messaging, and bulletin boards, to mention a few. This politeness is not just the words used, but also the software actions taken. The relative value of the proposed politeness sub-aspects (respect, openness, helpfulness, and remembering) can also be compared. Correlational studies could compare rated application politeness with market success. Longitudinal studies could determine if
successful applications become more polite over time. Ethnographic studies could explore how users perceive polite and impolite software.

The scope of online politeness also bears investigation. The definition implies that young or inexperienced users will tolerate impolite agents like Mr. Clippy more than experienced users. Also, it has been proposed that for interactions mandated by law, or other coerced acts, politeness will apply less. Other individual differences including gender, age, and culture, may also mediate the user reaction to impolite software. Cultural differences in polite computing raise highly complex issues of roles and social structures and may affect the boundary between what is required and what is polite.

CONCLUSION

Polite software asks before it allocates computer resources, openly declares itself and its acts, does not unnecessarily interrupt or draw attention to itself, offers understandable choices, and remembers past interactions. Conversely, impolite software acts without asking, does things secretly, interrupts unnecessarily, offers confusing choices, and has no recall of its past interactions with you.

If polite software attracts users, impolite software can drive them away. This implies a new type of IS error—social error. A program syntax error fails to support the needs of the computer technology. A software usability error fails to support the psychological needs of the computer user. However, a social error means the software fails to support the equally critical needs of human social interaction. While users misunderstand systems designed with poor usability, they understand impolite software all too well, and that is why they walk away from the interaction. Whether a system fails because the computer cannot run it, the user cannot run it, or the user will not run it, makes no difference. The end effect is still that the application does not run. A software social error gives the same outcome as a software crash or user failure. Indeed, social errors may be even worse, as it is in the nature of people to actively seek retribution against those who wrong others in social interactions.

A future is envisaged where software politeness is a critical requirement for socio-technical system success, especially where user willingness to participate counts. Polite computing could be taught in system design classes, along with other system requirements. A “politeness seal” could credit applications that give rather than take user choice. If physical society in general sees the value of politeness, online society should follow that lead. As software becomes not only useful and usable but also polite, the Internet may at last recognize its social potential.

ACKNOWLEDGMENT

Many thanks to Guy Kloss, Massey University, for his very useful comments and insights.

REFERENCES


Cooper, A. (1999). The inmates are running the asylum—Why high tech products drive us crazy and how to restore the sanity. USA.


This work was previously published in the Handbook of Conversation Design for Instructional Applications, edited by R. Luppicini, pp. 419-436, copyright 2008 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).
Chapter 8.11

Neo-Symbiosis: The Next Stage in the Evolution of Human Information Interaction

Douglas Griffith
General Dynamics Advanced Information Systems, USA

Frank L. Greitzer
Pacific Northwest National Laboratory, USA

ABSTRACT

The purpose of this article is to re-address the vision of human-computer symbiosis as originally expressed by J.C.R. Licklider nearly a half-century ago and to argue for the relevance of this vision to the field of cognitive informatics. We describe this vision, place it in some historical context relating to the evolution of human factors research, and observe that the field is now in the process of re-invigorating Licklider’s vision. A central concept of this vision is that humans need to be incorporated into computer architectures. We briefly assess the state of the technology within the context of contemporary theory and practice, and we describe what we regard as this emerging field of neo-symbiosis. Examples of neo-symbiosis are provided, but these are nascent examples and the potential of neo-symbiosis is yet to be realized. We offer some initial thoughts on requirements to define functionality of neo-symbiotic systems and discuss research challenges associated with their development and evaluation. Methodologies and metrics for assessing neo-symbiosis are discussed.

BACKGROUND

In 1960, J.C.R. Licklider wrote in his paper “Man-Machine Symbiosis,”

The hope is that in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today (p. 5).
This statement is breathtaking for its vision — especially considering the state of computer technology at that time, that is, large mainframes, punch cards, and batch processing. The purpose of this article is to re-address Licklider’s vision and build upon his ideas to inform contemporary theory and practice within the broader field of human factors as well as to offer a historical perspective for the emerging field of cognitive informatics.

It is curious to note that Licklider did not use the term symbiosis again, but he did introduce more visionary ideas in a symbiotic vein. A paper he co-authored with Robert Taylor, titled “The Computer As a Communication Device,” made the bold assertion, “In a few years, men will be able to communicate more effectively through a machine than face to face” (p. 21). Clearly the time estimate was optimistic, but the vision was noteworthy. Licklider and Taylor described the role of the computer in effective communication by introducing the concept of “On-Line Interactive Vicarious Expediter and Responder” (OLIVER), an acronym that by no coincidence was chosen to honor artificial intelligence researcher and the father of machine perception, Oliver Selfridge. OLIVER would be able to take notes when so directed, and would know what you do, what you read, what you buy and where you buy it. It would know your friends and acquaintances and would know who and what is important to you. This paper made heavy use of the concept of “mental models,” relatively new to the psychology of that day. The computer was conceived of as an active participant rather than as a passive communication device. Remember that when this paper was written, computers were large devices used by specialists. The age of personal computing was off in the future.

Born during World War II, the field of human factors engineering (HFE) gained prominence for its research on the placement of controls — commonly referred to as knobology within the field of HFE, which was an unjust characterization. Many important contributions were made to the design of aircraft, including controls and displays. With strong roots in research on human performance and human errors, the field gained prominence through the work of many leaders in the field who came out of the military: Alphonse Chapanis, a psychologist and a Lieutenant in the U.S. Air Force; Alexander Williams, a psychologist and naval aviator; Air Force Colonel Paul Fitts; and J.C.R. Licklider. Beginning with Chapanis, who realized that “pilot errors” were most often cockpit design errors that could be corrected by the application of human factors to display and controls, these early educators were instrumental in launching the discipline of aviation psychology and HFE that led to worldwide standards in the aviation industry. These men were influential in demonstrating that the military and aviation industry could benefit from research and expertise of the human factors academic community; their works (Fitts, 1951a) were inspirational in guiding research and design in engineering psychology for decades. Among the most influential early articles in the field that came out of this academic discipline was George Miller’s (1956) “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity to Process Information,” which heralded the field of cognitive science and application of quantitative approaches to the study of cognitive activity and performance.

An early focus of HFE was to design systems informed by known human information processing limitations and capabilities — systems that exploit our cognitive strengths and accommodate our weaknesses (inspired by the early ideas represented in the Fitts’ List that compared human and machine capabilities; Fitts, 1951b). While the early HFE practice emphasized improvements in the design of equipment to make up for human limitations (reflecting a tradition of machine centered computing), a new way of thinking about human factors was characterized by the design of the human-machine system, or more generally, human- or user-centered computing (Norman &
Draper, 1986). The new subdiscipline of interaction design emerged in the 1970s and 1980s that emphasizes the need to organize information in ways to help reduce clutter and “information overload” and to help cope with design challenges for next-generation systems that will be increasingly complex while being staffed with fewer people. Emphasis on human cognitive processes, and on the need to regard the human-machine system as a joint cognitive system, represented a further refinement that has been called cognitive systems engineering (Hollnagel & Woods, 1983).

Fundamental to all of these approaches and perspectives on HFE is the overriding principle to “know your user.” In a recent critical essay, Don Norman (2005) asks us to re-assess the human-centered design perspective: Developed to overcome the poor design of software products, human-centered design emphasized the needs and abilities of users and improved the usability and understandability of products. But despite these improvements, software complexity is still with us. Norman goes on to ask why so many designs of everyday things work so well, even without the benefit of user studies and human-centered design. He suggests that they all were “developed with a deep understanding of the activities that were to be performed (p.14).” Successful designs are those that fit gracefully into the requirements of the underlying activity. Norman does not reject human-centered design, but rather encompasses it within a broader perspective of activity-centered design. Further, he questions a basic tenet of human-centered design that technology should adapt to the human, rather than vice versa. He regards much of human behavior as an adaptation to the “powers and limitations of technology.” Activity-centered design aims to exploit this fact.

Other perspectives suggest that the focus of design should be on human-information interaction rather than human-computer interaction. Gershon (1995) coined the term Human-Information Interaction (HII) to focus attention on improving the way people “find, interact with, and understand information.” As such, HII includes aspects of many traditional research efforts, including usability evaluation methods and cognitive task analysis, but also design concepts that address the ethnographic and ecological environment in which action takes place. Examples of work in this area include distributed cognition (Zhang & Norman, 1994), naturalistic and recognition-primed decision making (Zsombok, 1997); and information foraging and information scent (Pirolli & Card, 1999).

In summary, over the last half century or so, the field of human factors has evolved through a series of modest perspective shifts and insights that have yielded a fair degree of success in approaches, methods, and techniques for design and evaluation of systems that are created to support and enhance human-information interaction. The many labels that have been applied to the field (cognitive engineering, human-centered computing, participatory design, decision centered design, etc.) are all “differently hued variants of the same variety” (Hoffman, Feltovich, Ford, Woods, Klein & Feltovich, 2002).

Engineering psychology and human factors are moving to a more encompassing scope of the field. Raja Parasuraman (2003) married neuroscience with ergonomics and termed it neuroergonomics. Don Norman (2004) incorporated affect (emotion) into the field with his book, Emotional Design: Why We Love (or Hate) Everyday Things. Hancock, Pepe and Murphy (2005) are developing the concept of hedonomics. They have developed a hierarchy of ergonomics and hedonic needs derived from Maslow’s (1970) hierarchy of needs: safety, the prevention of pain, forms the foundation of this pyramid; next comes functionality, the promulgation of process; then usability, the priority of preference (the transition from ergonomics to hedonomics begins at the usability layer); the next layer is pleasurable experience; and the apex of the pyramid comprises individuation and personal perfection. So the field is beginning to address the enhancement of individual potential.
Recent research in the emerging field of cognitive informatics (Wang, 2005a, b) addresses Maslow’s hierarchy of needs within a formal model that attempts to capture the relationships among human factors and basic human needs.

Recently a new research thrust has emerged that aims to shift the focus once more to not only enhancing the interaction environment, which is the aim of cognitive systems engineering, but also to enhance the cognitive abilities of the human operators and decision makers themselves. The Augmented Cognition program (Schmorrow & Kruse, 2004) within the DARPA Information Processing Technology Office (IPTO) aims to monitor and assess the user’s cognitive state through behaviorally and neurologically derived measures acquired from the user while interacting with the system and then to adapt or augment the computational interface to improve performance of the user-computer system. Schmorrow and McBride (2005) explain that this research is based on the view that the weak link in the human-computer system may be attributed to human information processing limitations, and that human and computer capabilities are increasingly reliant on each other to achieve maximal performance. Much of the research within the augmented cognition program seeks to further our understanding of how information processing works in the human mind so that augmentation schemes might be developed and exploited more effectively — in a variety of domains from clinical restoration of function to education to worker productivity to warfighting superiority. Thus, as described by Schmorrow and McBride: “the DARPA Augmented Cognition program at its core is an attempt to create a new frontier, not by optimizing the friendliness of connections between human and computer, but by reconceptualizing a true marriage of silicon- and carbon-based enterprises [International of Journal Human Computer Interaction, p. 128].”

While augmented cognition exploits neuroscience research as a path toward symbiosis of humans and machines, research in cognitive informatics embraces neuroscience research as a potential model and point of departure for “brain-like” machine-based cognitive systems that may someday exhibit human-like properties of sensation, perception, and other complex cognitive behaviour (Anderson, 2005a, b). We believe that neo-symbiosis provides a strong contextual framework to organize and guide research in cognitive informatics.

**NEO-SYMBIOSIS**

Once more, then, we are on the threshold of resurrecting a vision of symbiosis — but today we have the advantage of far greater computational resources and decades of evolution in the field of human factors/cognitive engineering. Licklider’s notion of symbiosis does require updating. First, the term “man/machine symbiosis” is politically incorrect and would be more appropriately termed “human/machine symbiosis.” Then there is a problem with the term symbiosis itself. Symbiosis implies co-equality between mutually supportive organisms. However, we contend that the human must be in the superordinate position. The Dreyfuses (Dreyfus, 1972, 1979, 1992; Dreyfus & Dreyfus, 1986) have made compelling arguments that there are fundamental limitations to what computers can accomplish, limitations that will never be overcome (Dreyfus & Dreyfus, 1986). In this case, it is important that the human remain in the superordinate position so that these computer limitations can be circumvented. On the other hand, Kurzweil has argued for the unlimited potential of computers (Kurzweil, 1999). But should it be proven that computers do, indeed, have this unlimited potential, then some attention needs to be paid to Bill Joy and his nightmarish vision of the future should technology go awry (Joy, 2000). In this case, humans would need to be in the superordinate position for their own survival.

Griffith (2005a) has suggested the term neo-symbiosis for this updated vision of symbiosis.
The augmented cognition research community is taking Licklider’s vision quite literally in exploring technologies for acquiring, measuring, and validating neurological cognitive state sensors to facilitate human-information interaction and decision-making. Neurobiologically inspired forms of symbiosis, while consistent with the metaphor that Licklider used, were not a focus of Licklider’s vision; but the possibilities for enhanced cognitive performance are enticing. Clearly, however, much work is required to achieve a brain-computer interface that might be called neo-symbiotic. Much of the effort in this field to date has focused on cognitive activity that tends to be more oriented toward attention and perception processes, and less toward decision making and thinking. In this sense, augmented-cognition neurological inputs can help to approach neo-symbiosis by providing information to the computer that can in turn be fed back to the human in the form of adaptive displays and interactions or other functions aimed to mitigate the effects of stress or information overload. More ambitious goals of increasing total cognitive capacity through augmented cognition technologies are still on the horizon of this research program and recent offshoots of augmented cognition R&D such as DARPA’s Neurotechnology for Intelligence Analysts program. Our interest, similarly, is in the current potential for enhanced human-computer collaboration that will achieve a level of performance that is superior to either the human or the computer acting alone.

The principal reason that the beginning of the 21st century is so propitious for the reinvigoration of Licklider’s vision is the result of advancements in computer technology and psychological theory. Therefore, one of our major objectives is to increase the human’s understanding, accuracy, and effectiveness by supporting the development of creative insights. Understanding involves learning about the problem area and increasing the variety of contexts from which the problem can be understood. Enhanced accuracy/effectiveness can be achieved by endowing the computer with a variety of means to support the task or activity. Revisiting thoughtful prescriptions for such computer-based intelligent support capabilities from two decades ago, we find examples such as knowledge of the user’s goals and intentions, contextual knowledge (Croft, 1984), and “cognitive coupling” (Fitter & Sime, 1980) functions that include (Greitzer, Hershman & Kaiwi, 1985) the ability to inform the user about the status of tasks, remind the user to perform certain tasks, advise the user in selecting alternative actions, monitor progress toward the goal, anticipate requests to display or process information, and test hypotheses. In the context of information analysis tasks, examples of such neo-symbiotic contributions by the computer include considering alternative hypotheses, assessing the accuracy of intelligence sources, and increasing the precision of probability estimates through systematic revision. These types of activity-based support functions, enhanced by cognitive models, are the concepts that we believe will put us more solidly on the path to the original vision of Licklider, a neo-symbiosis where there is a greater focus on cognitive coupling between the human user and the computer.

**NEO-SYMBIOSIS RESEARCH AGENDA**

**Requirements: Implementing Neo-Symbiosis**

How should neo-symbiosis be implemented? Fortunately, Kahneman (2002, 2003) and Kahneman and Frederick (2002) has provided guidance through a theoretical framework. In his effort to organize seemingly contradictory results in studies of judgment under uncertainty, he has advanced the notion of two cognitive systems introduced by Sloman (1996, 2002) and others (Stanovich, 1999; Stanovich & West, 2002). System 1, termed Intuition, is fast, parallel, automatic, effortless,
associative, slow learning, and emotional. System 2, termed Reasoning, is slow, serial, controlled, effortful, rule-governed, flexible, and neutral. The cognitive illusions, which were part of the work for which he won the Nobel Prize, as well as perceptual illusions, are the results of System 1 processing. Expertise is primarily a resident of System 1. So are most of our skilled performance such as recognition, speaking, driving, and many social interactions. System 2, on the other hand, consists of conscious operations, such as what is commonly thought of as thinking. Table 1 summarizes these characteristics and relationships. The upper portion of the table describes human information processing characteristics and strengths, interpreted within Kahneman’s (2003) System 1/System 2 conceptualization. The bottom portion of the table represents an update of traditional characterizations of functional allocation based on human and computer capabilities, such as the original Fitts’ List (Fitts, 1951b), cast within the System 1/System 2 framework.

System 1 is effective presumably due to evolutionary forces, massive experience, and by constraining context. Most of the time, it is quite effective. System 1 uses nonconscious heuristics to achieve these efficiencies, so occasionally it errs and misfires. Such misfires are responsible for perceptual and cognitive errors. One of the roles of System 2 is to monitor the outputs of System 1 processes. It is the System 2 processes that require computer support, not only with respect to the pure drudgery and slowness of human System 2 processes, but also with respect to the monitoring of System 1 processes. In most cases, however, it is a mistake to assign System 1 processes to the computer. This was the fundamental error in many automatic target recognition and image interpretation algorithms that attempted to automate the human out of the loop. Even to this day, computer technology has been unsuccessful in modeling human expertise in System 1 domains. The perceptual recognition processes of most humans are excellent. System design should capitalize upon these superb processes and provide support to other areas of human information processing such as search (there is a tendency to overlook targets); interpretation keys to provide a check and support for the recognition process; analysis and synthesis (e.g., to augment reasoning processes); support to facilitate adjusting to changes in context (e.g., to maintain situational awareness); and computational support (e.g., to make predictions). The bottom portion of Table 1 exhibits examples of how human and computer contributions can be allocated to System 1 and System 2 processing in a neo-symbiotic system.

Greitzer (2005b) has discussed the importance of identifying cognitive states in real-world decision-making tasks. A critical question here is, what are the cognitive states that need to be measured? What are the cognitive states that, if identified and measured, could enhance neo-symbiosis? Clearly it would be beneficial to identify neurological correlates for System 1 and System 2 processes. It would be especially beneficial to identify neurological correlates of System 2 while monitoring System 1 processing. Perhaps there is a neurological signature when potential errors are detected in System 1 processing. It is conceivable that some of these errors remain below the threshold of consciousness. If these errors were detectable in the neurological stream, computers could assist in this error monitoring process.

As was mentioned previously, the identification of neurological correlates is not a requirement, nor is it the only enabler for neo-symbiosis. Griffith (2005b) has argued that neo-symbiosis can be achieved over a wide range of technological sophistication. Overviews and tutorials can be presented on basic human information processing capabilities, limitations, and biases. A software agent, or avatar, can pop up at strategic times with reminding prompts or checklists. Of course, the capability to monitor the human’s cognitive state through neurological correlates will enhance the ability of the avatar to pop up at strategic times. It might also be possible to monitor the content
# Neo-Symbiosis

## Table 1. System 1 and System 2 processes

<table>
<thead>
<tr>
<th>Human Processes</th>
<th>System 1: Intuition</th>
<th>System 2: Reasoning</th>
</tr>
</thead>
</table>
| **Processing Characteristics** | o Fast  
 o Parallel  
 o Automatic  
 o Effortless  
 o Associative  
 o Slow-Learning  
 o Emotional | o Slow  
 o Serial  
 o Controlled  
 o Effortful  
 o Rule-governed  
 o Flexible  
 o Neutral |
| **Type of Processing** | o Expertise  
 o Skilled Performance  
 o Most Perception | o Thinking  
 o Goal-driven Performance  
 o Anomaly and Paradox Detection |

<table>
<thead>
<tr>
<th>Neo-Symbiotic Functions</th>
<th>System 1: Intuition</th>
<th>System 2: Reasoning</th>
</tr>
</thead>
</table>
| **Examples of Human Contributions** | o Providing Context  
 o Detecting Contextual Shifts  
 o Intuition  
 o Pattern Recognition  
 o Creative Insights | o Supervision/Monitoring  
 o Inductive Reasoning  
 o Adaptability to Change  
 o Contextual Evaluations  
 o Anomaly Recognition/Detection  
 o Goal-Driven Processes/Planning  
 o Creative Insights |
| **Examples of Computer Contributions** | o Recognize Cognitive State Changes  
 o Adapt Displays/Interaction Characteristics to Human’s Cognitive State | o Deductive Reasoning  
 o Search  
 o Situational Awareness  
 o Analysis/Synthesis  
 o Hypothesis Generation/Tracking  
 o Computational Support  
 o Information Storage/Retrieval  
 o Multiprocessing  
 o Update Status of Tasks  
 o Advise on Alternatives  
 o Monitor Progress  
 o Monitoring System 1 Processes |

* This portion of the table based on Kahneman (2003)
of the interactions with the computer to identify potential processing problems. Differences in processing time present is yet another potential source of information for detecting errors and biases.

In our view, the thrust of the HII research agenda should be targeted at enhancing neo-symbiosis. A major focus of HII research today is aimed at visualization technology that processes and seeks to represent massive data in ways that facilitate insight and decision making. Data visualization technology seeks to facilitate visual thinking to gain an understanding of complex information, and perhaps most particularly to gain insights that would otherwise not be apparent from other data representations. A famous example of a successful visualization is the periodic table of elements (conceived by Mendeleev and published in 1869), which not only provided a simple display of known data but also pointed out gaps in knowledge that led to discoveries of new elements. However, creating novel visualizations of complex data (information visualization) does not guarantee success; there are arguably more examples of visualizations that have not lived up to expectations than success stories. A leap of faith is required to expect that a given scientific visualization will produce the “aha!” moment that leads to an insightful solution to a difficult problem. We assert that the key to a successful scientific visualization is its effectiveness in fostering new ways of thinking about a problem — in the System 1 sense as exemplified in Table 1 (e.g., seeing contextual shifts, recognizing new patterns, finding creative insights). This view stresses that the interaction component of HII needs to be emphasized. The human should not be regarded as simply a passive recipient of information display, however creative that information display might be. The human needs to be able to manipulate and interact with the information. The ability to manipulate information and view it in different contexts is key to the elimination of cognitive biases and to the achievement of novel insights (e.g., finding the novel intelligence in massive data). The goal is a neo-symbiotic interaction between the human and the information.

Thus, requirements should be defined so that a neo-symbiosis can be achieved between humans and their technology. Questions to guide the requirements definition process for neo-symbiotic systems designed to facilitate HII include:

- **How can such systems be designed to mitigate or eliminate cognitive biases?** Detecting/recognizing possible bias is one part of the challenge; an equally critical R&D goal is to define mitigation strategies. What types of interventions will be effective, and how should interventions be managed? We suggest that a mixed-initiative solution will be required that maintains the supervisory control of the human.

- **How can such systems be designed to leverage the unique processing skills of humans?** A prerequisite here is to identify the unique processing skills of humans. Technologies and approaches for developing idiosyncratic user models would be most useful. Moreover, expert users can identify and contribute their own unique skills: Consider an image interpretation system in which an expert with knowledge of a certain area could correct and elaborate upon outputs of image interpretation algorithms.

- **How can such systems be designed to facilitate collaboration?** One aim is to realize the assertion made by Licklider and Taylor (1968) that people will be able to communicate more effectively through a machine than face to face.

- **How can such systems promote a more pleasurable experience?** The goal here is to address some of the objectives outlined by Hancock et al. (2005).

- **How can such systems help someone to leverage personal potential or overcome a personal deficit (e.g., through augmenta-
Neo-Symbiosis

tive/assistive technology? A major area of interest for neurally-based symbiotic studies is the use of implant technology in which a connection is made between technology and the human brain or nervous system. Important medical applications include restoring lost functionality in individuals due to neurological trauma or a debilitating disease, or for ameliorating symptoms of physical impairment such as blindness or deafness. Other applications that do not address medical needs but instead aim to enhance or augment mental or physical attributes provide a rich area of research in the growing area of augmented cognition. Warwick and Gasson (2005) review the field of research and describe his research and experiences as a researcher and experimental subject who is the first human to have a computer chip inserted into his body that enabled bidirectional information flow and demonstration of control of a remote robot hand using the subjects’ own neural signals (Gasson, Hutt, Goodhew, Kyberd & Warwick, 2002; Warwick & Gasson, 2005). Warwick and Gasson (2005) observe:

By linking the mental functioning of a human and a machine network, a hybrid identity is created. When the human nervous system is connected directly with technology, this not only affects the nature of an individual’s ... identity, ... but also it raises serious questions as to that individual’s autonomy.

It should be appreciated, however, that assistive technology need not necessarily entail implants or any involvement with neurology. Indeed a great deal has already been accomplished via adaptive software and input and output devices (Griffith, 1990; Griffith, Gardner-Bonneau, Edwards, Elkind & Williges, 1989).

What are implications and requirements for computer architectures to achieve neo-symbiosis? A central point underlying neo-symbiosis is that humans need to be included in the computer architecture or system design. It is anathema to the concept of neo-symbiosis that computers and humans be regarded in isolation. They need to be considered together with the objective of each exploiting the other’s potential and compensating for the other’s weaknesses. Ideally the interaction between the two will achieve a multiplicative effect, a true leveraging.

Metrics: Measuring Success

An important question is how to identify neo-symbiotic design and how to assess it. It is important to recognize instances of neo-symbiotic design that are already among us in the form of productivity enhancement tools or job aids. For example, spell checking in contemporary word processors compensate for memory and perceptual/motor shortcomings; thesauruses leverage communicative abilities. Various creativity tools, such as concept mapping, leverage creative potential. In the augmented cognition domain, various neurologically-based “cognitive state sensors” are emerging as indicators of cognitive load and as potential cognitive prosthetics for medical purposes. In each of these cases, particularly the most recent developments that aim to enhance cognitive functions and effectiveness, evaluation methods and metrics are needed to guide research and facilitate deployment of technologies. For more advanced development of neo-symbiotic designs that aim to enhance human information processing and decision making (e.g., intelligence analysis performance) or knowledge/skill acquisition (e.g., training applications), we recognize the need for more rigorous evaluation methods and
metrics that reflect the impact of the technology on performance.

Of course, standard subjective measures can readily be expanded to include neo-symbiotic potential. Many subjective measures are interpreted in terms of usability. There are several sources of established guidelines for usability testing (e.g., Nielsen, 1993). Commonly used criteria include efficiency, ability to learn, and memorability. Usability measures the address of the experience of users; whether or not they found the tool useful, easy to learn, easy to use, and so forth. Often, users are asked to provide this sort of feedback using qualitative measures obtained through verbal (“out loud”) protocols and/or post-hoc comments (via questionnaires, interviews, ratings). Likert scales, in which respondents indicate their degree of agreement or disagreement with particular statements using numerical ratings, can use question stems such as: “Using this application/system enhanced my performance”; or “Using this application/system compensated for my information processing shortcomings.”

Subjective measures such as these are designed to assess the acceptance by users of the system. It is unfortunate that the term subjective is used in a pejorative sense and that subjective measures are all too often regarded as second rate measures. Whether or not a system is perceived favorably and judged to be useful are central questions in evaluating the system’s value. Especially relevant to neo-symbiosis is the user’s assessment of the extent to which his or her potential has been enhanced.

It is possible to use magnitude estimation to assess the subjective amount, or lack of, neo-symbiosis in an application/system. In magnitude estimation (Stevens, 1975), stimuli are evaluated with respect to a standard stimulus, or modulus. That standard stimulus is assigned a value, and other stimuli are evaluated proportionate to it. So if the modulus was assigned a value of 50, and the stimulus being rated was regarded as half of whatever the rating dimension was, it would be rated 25. Were it regarded as having twice the value on the rating dimensions, it would be rated 100. A given version of Microsoft Word™ could be assigned a value of 50. If someone regarded another word processor as being twice as neo-symbiotic as this version of Word, it would be rated 100. Were it regarded to be only half as neo-symbiotic, it would be rated 25. A desirable property of magnitude estimation methods is that they produce ratio scales. Magnitude estimation is a remarkably robust methodology. Its validity has been demonstrated with stimuli ranging from the loudness of tones to the seriousness of crimes. It uses an anchor to a standard that allows proportional assessments of where an issue, item, stimulus stands with respect to that standard. Thus, statements can be made that a product is 20% better than a related product, 40% worse, and so forth. These ratings are more meaningful and interpretable than many other subjective rating techniques.

Whenever feasible, subjective measures should be supplemented with objective measures. Greitzer (2005a) has argued for development of measures of effectiveness based on performance impact in addition to the continued use of traditional subjective usability measures. User satisfaction is a necessary, but not sufficient measure. Behavioral measures are needed to address more cognitive factors and the utility of tools or technologies: Does technology X improve the throughput of cognitive tasks of type Y? Does it yield more efficient or higher quality output for certain types of tasks? Quantitative measures that assess utility may include efficiency in completing the task (time, accuracy, completeness). These will be most useful in comparing the utility of alternative tools or assessing the utility of a given tool vs. baseline performance without the tool. For example, in information analysis tasks, it has been observed (Scholtz, Morse & Hewett, 2004) that analysts tend to spend more time in data collection and report generation than in analysis activity (hence a kind of “bathtub curve” as described by Wilkins
of the work in human factors and ergonomics has been empirical. Only occasionally has the field drawn upon theory. The field of HII has been primarily technology driven. Programs and systems are developed on the bases of intuitions and what is regarded as cool and challenging by the developer, rather than from considerations of the information processing shortcomings and potential of the users. Very often techniques are not even subject to empirical assessment. But a strategy of generating an idea and then evaluating it empirically will not prove successful in the long run. HII requirements need to be developed not only on the basis of what a given system is being designed to accomplish, but also on the basis of theory and data in cognitive science and neuroscience.

To sum up, we have argued that the field of HII is on the threshold of realizing a new vision of symbiosis — one that embraces the concept of mutually supportive systems, but with the human in a leadership position, and that exploits the advances in computational technology and the field of human factors/cognitive engineering to yield a level of human-machine collaboration and communication that was envisioned by Licklider, yet not attained. As we have described, the field of human factors/HII is not static, but rather must inexorably advance. With advances in computer technology, cognitive science, and neuroscience, human potential and fulfillment can be leveraged more, yielding a spiral of progress: As human potential is raised, then that new potential can be leveraged even further. We think this vision provides a useful framework for cognitive informatics.

REFERENCES


**ENDNOTES**

1 A research program at DARPA, Neurotechnology for Intelligence Analysts, seeks to identify robust brain signals that may be recorded in an operational environment and that are correlated with imagery data of potential interest to the analyst. Investigations of visual neuroscience mechanisms have indicated that the human brain is capable of responding to visually salient objects significantly faster than an individual’s visuomotor response—i.e., essentially before the human indicates awareness. The program seeks to develop information processing triage methods to increase the speed and
Neo-Symbiosis


As Anderson (2005b) has observed, human expertise in System I domains has been very difficult to model in computers, and many researchers (connectionists, behavior-based roboticists) have used this to argue that digital computer metaphor is flawed.

This work was previously published in the International Journal of Cognitive Informatics and Natural Intelligence, edited by Y. Wang, Volume 1, Issue 1, pp. 39-52, copyright 2007 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).
Emerging pervasive computing scenarios involve client applications that dynamically collect information directly from the local environment. The sophisticated distribution and dynamics involved in these applications place an increased burden on developers that create applications for these environments. The heightened desire for rapid deployment of a wide variety of pervasive computing applications demands a new approach to application development in which domain experts with minimal programming expertise are empowered to rapidly construct and deploy domain-specific applications. This chapter introduces the DAIS (Declarative Applications in Immersive Sensor networks) middleware that abstracts a heterogeneous and dynamic pervasive computing environment into intuitive and accessible programming constructs. At the programming interface level, this requires exposing some aspects of the physical world to the developer, and DAIS accomplishes this through a suite of novel programming abstractions that enable on-demand access to dynamic local data sources. A fundamental component of the model is a hierarchical view of pervasive computing middleware that allows devices with differing capabilities to support differing amounts of functionality. This chapter reports on our design of the DAIS middleware and highlights the abstractions, the programming interface, and the reification of the middleware on a heterogeneous combination of client devices and resource-constrained sensors.
INTRODUCTION

As networked computing capabilities become increasingly ubiquitous, we envision an instrumented environment that can provide varying amounts of information to applications supporting mobile users immersed within the network. While such a scenario relies on low-cost, low-power miniature sensors, it deviates from existing deployments of sensor networks, which are highly application-specific and generally funnel information to a central collection service for a single purpose. Instead, solutions for ubiquitous computing must target future scenarios in which multiple mobile applications leverage networked nodes opportunistically and unpredictably. To date, most application development for ubiquitous computing has been limited to academic circles. One significant barrier to the widespread development of ubiquitous computing applications lies in the increased complexity of the programming task when compared to existing distributed or even mobile situations. Sensor nodes, which provide computational platforms embedded in the environment, are severely resource-constrained, in terms of both computational capabilities and battery power, and therefore, application development must inherently consider low-level design concerns. This complexity, coupled with the increasing demand for ubiquitous applications, highlights the need for programming platforms (i.e., middleware) that simplify application development.

As will be described in more detail in later sections, much existing work in simplifying programming in sensor networks focuses on application-specific networks where the nodes are statically deployed for a particular task. Ubiquitous computing requires a more futuristic (but not unrealistic) scenario in which sensor networks become more general-purpose and reusable. While the networks may remain domain-specific, ubiquitous computing applications that will be deployed are not known \textit{a priori} and may demand varying capabilities from the environment. Finally, existing applications commonly assume that sensor data is collected at a central location to be processed and used in the future and/or accessed via the Internet. Applications for ubiquitous computing, however, involve users immersed in a network environment who access locally sensed information on demand. This is exactly the vision of pervasive computing environments (Weiser, 1991), in which sensor networks must play an integral role (Estrin, Culler, Pister, & Sukhatme, 2002).

While this style of interaction is common in many application domains, we will refer to applications from the first responder domain, which provides a unique and heterogeneous mix of embedded and mobile devices. The former includes fixed sensors in buildings and environments that are present regardless of crises and ad hoc deployments of sensors that responders may distribute when they arrive. Mobile devices include those moving within vehicles, carried by responders, and even autonomous robots that may perform exploration and reconnaissance.

In this chapter, we will first demonstrate that immersive networks built of tiny sensing and actuating devices are an essential component of the vision of ubiquitous computing. We will carefully define what we mean by immersive networks and highlight how the definition differs from traditional uses of such technologies. Because these networks deviate so significantly from existing application environments, constructing tailored, adaptive applications necessary for ubiquitous computing is difficult. We will demonstrate this complexity through a handful of example applications and then provide a detailed survey of existing approaches to simplifying the development of ubiquitous computing applications that rely on immersive sensor networks. Following this survey of existing technologies, we will introduce a new paradigm for programming ubiquitous immersive networks that not only simplifies the programming task, but also improves the performance of
enabling programmable ubiquitous computing environments

the supportive network in terms of battery life, communication overhead, and latency.

the next section describes the operating environment of immersive sensor networks for supporting ubiquitous computing and provides a comparative analysis of existing approaches to simplifying programming for ubiquitous computing. We then introduce a new programming paradigm for immersive sensor networks that combines communication abstractions with high-level programming constructs.

background

in ubiquitous computing applications that rely on immersive sensor networks, users with client devices need to interact directly with devices (or sensors) embedded in their environments. this allows the client applications to operate over information collected directly from the local area (as shown in Figure 1(b)). this is in contrast to existing sensor network deployments in which sensor networks are commonly accessed through a central collection point (as shown in Figure 1(a)). More directly, immersive sensor networks support ubiquitous computing applications, not remote distributed sensing. in this type of environment, we differentiate client devices (those on which ubiquitous computing applications run) from sensors (devices embedded in the environment). the former commonly support users and have increased computational power, while the latter are heavily resource-constrained.

Figure 1. Comparison of (a) existing operational environments and (b) immersive sensor networks for ubiquitous computing

(a) (b)
Supporting Ubiquitous Computing: An Operating Environment

The style of interaction apparent in ubiquitous computing and depicted in Figure 1(b) differs from common uses of sensor networks. First and foremost, applications require direct local interactions with embedded devices present in the immediately accessible environment. Second, the area from which an application desires to draw information is subject to the device user’s movement. These mobility-induced dynamics demand a constant reevaluation of the sensors participating in ongoing interactions. Third, networks must support general classes of applications. Few a priori assumptions can be made about the needs or intentions of applications, requiring a middleware to handle unpredictable coordination. Finally, in immersive sensor networks, the programming complexity is drastically increased due to the above concerns coupled with traditional concerns associated with distributed, embedded computing. In addition, the desire to provide end-user applications (as opposed to more database-oriented data collection) increases the demand for applications and the number of programmers that will need to construct them.

Ubiquitous computing applications from domains such as intelligent construction sites (Hammer et al., 2006), aware homes (Kidd et al., 1999), and pervasive office environments (Voida, Maynatt, & MacIntyre, 2002) involve users immersed in the network who access locally sensed information on demand. Such application scenarios motivate a view of the ubiquitous computing environment that is device-agnostic. That is, applications, in general, do not care about devices embedded in the environment, but instead about information and resources available locally. Logically, the ubiquitous environment appears as a world of embedded information available for immersed applications. To support this view, it is essential to allow developers to distance themselves from explicit knowledge of devices and communication capabilities of the underlying pervasive computing network and instead focus on dynamically changing, locally available information.

Sensor networks for these environments will also need to provide reusability. Current efforts offer solutions that are specific to a particular application (e.g., once the sensor network has been tasked to do habitat monitoring, that is all it is expected to do), but the future will see multipurpose networks deployed to support numerous applications whose natures may not be known at the time the network is deployed. The cost of physically visiting each sensor to program it is prohibitive, and therefore, the ability to remotely and dynamically tailor sensor networks to particular applications will be essential.

The data collection schemes used in existing deployments of sensor networks commonly require sensors to relay raw data to sink nodes to perform further processing. For the sensing devices, communication is much more expensive than local computation; thus, the approaches used in existing deployments lead to short network lifetimes. Furthermore, the throughput at each node decreases as the network scales due to redundant broadcasts, leading to inefficient use of the network bandwidth. To support ubiquitous environments, sensor networks will need to perform complicated tasks and in-network processing to seamlessly transform raw data.

Related Work: Simplifying Programming for Ubiquitous Computing

It is largely recognized that constructing applications for ubiquitous computing environments is a significant undertaking. Several approaches have made strides in simplifying the kind of programming necessary for immersive networks. This section provides a thorough comparative analysis of these approaches.
investigation into these existing techniques, from middleware solutions to toolkits and programming languages.

**Ubiquitous Computing Middleware**

One strong example of a middleware for ubiquitous computing, Gaia (Roman et al., 2002), introduces *active spaces* as a programmable environment, by encapsulating the heterogeneity of devices that are located in them. It abstracts user data and applications into a *user virtual space* that has a dynamic mapping to the resources in the current environment. Users always have their virtual space available, even as they move across different active spaces. Furthermore, they can simultaneously interact with multiple devices, dynamically reconfigure applications, pause and resume applications, and use context attributes to program application behaviors (Roman et al., 2002). However, this model assumes a centralized system structure which is in direct opposition to the goal of deploying large numbers of applications over a widely dispersed immersive sensor networks.

**Blackbox Abstractions**

In contrast, projects targeted directly for sensor networks more directly address the desire to reduce computational and power requirements and to operate in a more distributed fashion. Two demonstrative examples that have explored representing the sensor network as a database are TinyDB (Madden, Franklin, Hellerstein, & Hong, 2005) and Cougar (Yao & Gehrke, 2002). Generally, these approaches enable applications to generate data requests that flow out from a central point (i.e., a base station) and create routing trees that funnel replies back to this root (in a manner similar to that shown in Figure 1(a)). Much of the work in these approaches focuses on performing intelligent in-network aggregation and routing to reduce the overall energy cost while still keeping the semantic value of data high. Each node processes streams of sensor data much like the processing of streams in a database. In both approaches, data aggregation is specified using an SQL-like language over homogeneous data types. Ubiquitous computing applications often require the use of many nearby sensors, ultimately aggregating these disparate pieces of data into a cohesive piece of information for the application or user. Therefore, despite the fact that moving data across the network in approaches such as TinyDB and Cougar still requires centralized algorithms, they have much to offer in support of ubiquitous computing applications in immersive sensor networks. REED (Abadi, Madden, & Lindner, 2005) is an extension to TinyDB that supports joins between sensor data and static tables built outside the sensor network. Users can express queries with complex time- and location-varying predicates over multiple conditions using join predicates over these different attributes. REED organizes nodes into groups and stores a part of the predicate table at each group member, reducing storage costs at remote nodes. This in-network application of *join* can significantly reduce the communication overhead in the network. However, since the group members contain parts of an external table, they are all required to be in radio range of one another.

State-centric programming (Liu, Chu, Liu, Reich, & Zhao, 2003) mediates between an application developer’s model of physical phenomena and the distributed execution of sensor network applications. It uses the notion of collaboration of groups to abstract common patterns in application-specific communication and resource allocation. Furthermore, it takes a signal processing and control theory approach, where dynamically created collaboration groups provide the input to applications written by application developers as algorithms for state update and retrieval. Consequently, the resulting software is more modular.
Enabling Programmable Ubiquitous Computing Environments

and can be ported to different platforms as these programs are less affected by system configuration changes. However, the state-centric programming implementation and evaluation relies on the software environment, PIECES (developed in Java and Matlab), which, unlike most other network simulators, does not simulate network behavior all the way down to the packet level (it only verifies the algorithms at the collaboration group level). While the implementation of collaboration groups needs to be distributed in real-world deployments, collaborations groups are implemented as centralized objects in PIECES. The Abstract Task Graph (ATaG) (Bakshi, Prasanna, Reich, & Larner, 2005) methodology provides system-level support for architecture-independent sensing application development. ATaG uses a combination of imperative and declarative programming styles and has a data-driven program flow. For example, in an environment monitoring application, it allows the periodic computation and logging of the maximum pressure in the system, and the periodic monitoring of temperature. However, it cannot combine these two different types of data to arrive at an abstracted measurement.

Virtual Machines and Code Generation

Other approaches have focused more specifically on the programmability of ubiquitous computing environments. VM* (Koshy & Pandey, 2005) is a virtual machine approach that can scale software components depending on the constraints of each device. This allows application developers to better manipulate unpredictable environments with a wide variety of devices, but has limitations in that the virtual machine must know about the applications in advance to be able to optimize resource usage. TinyGALS (Cheong, Liebman, Liu, & Zhao, 2003) allows programmers to represent applications in terms of relatively high-level components, which are subsequently synthesized into the low-level, lightweight, efficient programs that are deployed on the nodes. This eases the programming task but does not allow arbitrary applications to access the immersive sensor network and immediately start to use it. MiLAN (Heinzelman, Murphy, Carvallo, & Perillo, 2004) aims to enable applications to control the network’s resource usage and allocation optimally to tune the performance of an entire sensor network through the definition of application policies that are enacted on the network. MiLAN tries to maximize network lifetime as well as meet the application’s quality-of-service requirements. While such approaches are highly beneficial when the application is known and the networks are relatively application-specific, they do not map well to immersive sensor networks where the nodes must be able to service a variety of unpredictable applications.

Toolkits and Development Suites

More generalized approaches attempt to provide integrated suites of tools that enable simplified programming of sensor networks. For example, EmStar (Girod et al., 2004) provides a suite of libraries, development tools, and application services that focus on coordinating microservers (e.g., sensing devices with computational power equivalent to a PDA). However, EmStar functions only on Linux-based platforms such as the Stargate. The Sensor Network Application Construction Kit (SNACK) (Greenstein, Kohler, & Estrin, 2004) consists of a set of libraries and a compiler that makes it possible to write very simple application descriptions that specify sophisticated behavior using components written in nesC (the TinyOS programming language that runs on sensor motes). While EmStar and SNACK are programming environments for individual nodes, Agilla (Fok, Roman, & Lu, 2005) is an agent-based middleware that allows applications to inject agents into the sensor network that coordinate through local tuple spaces and migrate intelligently to carry out the applications’ tasks. Multiple autonomous applications
can run simultaneously over the sensor network. However, Agilla does not use any mechanisms for authenticating agent activities.

**Coordination Approaches**

One approach that does map well to the operational picture shown in Figure 1(b) is TinyLIME (Curino et al., 2005), a tuple space based middleware that enables mobile computing devices to interact with sensor data in a manner decoupled in both space and time. Applications create tuple templates to subscribe for data that is of interest to them. The tuple spaces of a pair of devices are temporarily federated whenever the devices are within a single hop of one another. TinyLIME allows client devices to connect to sensors available in the immediate environment, but does not enable multihop communication or aggregation. Another adaptation of the LIME model, TeenyLIME (Costa, Mottola, Murphy, & Picco, 2006), uses the abstraction of a shared tuple space that contains the data of the local device and its one-hop neighbors. Since TinyLIME targets sensor networks where users with mobile devices request data from sensors immediately around them, the applications are deployed on the client devices and the sensing devices are only data producers without any tuple spaces. On the other hand, TeenyLIME applications are deployed directly on the sensing devices that have their own tuple spaces and play an active role in distributed coordination.

**Location-Dependent Approaches**

EnviroTrack (Abdelzaher et al., 2004) is an object-based and data-centric middleware designed specifically for embedded tracking applications. EnviroTrack associates a context-label with each entity. Upon initial detection of the entity, the context-label is dynamically created and logically follows the entity’s movement through the sensor field. Application developers directly interact with the context label instead of a continuously changing collection of nodes that detect the entity, through the help of a directory service based on a geographic hash table. EnviroTrack relies on embedded sensors with precise knowledge of their locations to locate and track mobile objects.

**Macroprogramming Approaches**

Finally, Kairos (Gummadi, Gnamiali, & Govindan, 2005) is a macroprogramming model that allows the specification of the network’s global behavior through a centralized model. As such, it is not adaptive or general-purpose, requiring deployment-time knowledge of the intended application(s). Regiment (Newton & Welsh, 2004) also employs a macroprogramming approach to program sensor networks. A user writes a single program that is then distributed and run across the sensor network. Kairos provides abstractions to facilitate this task, while Regiment focuses on the suitability of functional programming to the sensor network domain.

In summary, while these systems for ubiquitous computing have addressed components of the problems associated with the operating environment described above, other components of the problem definition are not completely satisfied by these existing systems. Overall, the projects focus on specific facets of enabling application development in ubiquitous computing or sensor networks, but none take a holistic approach. We claim that any middleware for simplifying the creation of ubiquitous computing applications must provide constructs tailored to this unique and dynamic environment. These constructs must consider both communication and coordination aspects of tying together embedded devices as well as the application-level operations that are available to instruct these dynamic networks. In the next section, we introduce a new paradigm for immersive sensor networks targeted directly towards the application of such networks to the challenges posed by ubiquitous computing applications and their operating environments.
Enabling Programmable Ubiquitous Computing Environments

A NEW MIDDLEWARE PARADIGM FOR IMMERSIVE NETWORKS

In this section, we introduce a new programming paradigm for immersive sensor networks that combines communication abstractions with high-level programming constructs. We argue that such knowledge about the communication environment by the programmer is essential to mediating performance and efficiency concerns, and our approach minimizes the complexity of knowing about communication using a pair of intuitive grouping abstractions, the scene, and the virtual sensor.

Scenes: Abstractions of Local Data

In an immersive sensor network, a user’s operational context is highly dynamic. As the user moves through the environment, the set of embedded devices he interacts with should change accordingly. Furthermore, if the sensor network is well connected, the client device will be able to reach vast amounts of raw information that must be filtered to be usable. To enable efficient solutions, the application must be able to limit the scope of its interactions to include only the data that matches its needs. Therefore, an application developer must be conscious of the communication tasks required to satisfy his desired application behaviors.

In the middleware paradigm this chapter describes, such specification of the operating context is encapsulated in an abstraction called a scene (Kabadayi & Julien, 2007). Applications define scenes according to their needs, and each scene constrains which particular sensors may influence the application. The constraints may be on properties of hosts (e.g., battery life), of network links (e.g., bandwidth), and of data (e.g., type).

The declarative specification defining a scene allows an application programmer to flexibly describe the type of scene he wants to create. Multiple constraints can be used to define a single scene. The programmer only needs to specify three parameters to define a constraint:

- **Metric:** A property of the network or environment that defines the cost of a connection (i.e., a property of hosts, links, or data)
- **Path cost function:** A function (such as sum, average, minimum, maximum) that operates on a network path to calculate the cost of the path
- **Threshold:** The value a path’s cost must satisfy for that sensor to be a member of the scene.

Thus, a scene, $S$, is specified by one or more constraints, $C_1, C_2, \ldots, C_n$:

$$C_i = \langle M_i, F_i, T_i \rangle, \quad C_2 = \langle M_2, F_2, T_2 \rangle, \ldots, \quad C_n = \langle M_n, F_n, T_n \rangle$$

where $M$ denotes a metric, $F$ denotes a path cost function, and $T$ denotes a threshold.

Figure 2 demonstrates the relationships between these components. This figure is a simplification that shows only a one-constraint scene and a single network path. The cost to a particular node in the path (e.g., node $i$) is calculated by applying the scene’s path cost function to the metric. The metric can combine information about the path so far ($p_i$) and information about this node. Nodes along a path continue to be included in the scene until the path hits a node whose cost (e.g., $c_k$) is greater than the scene’s threshold. This functionality is implemented in a dynamic distributed algorithm that can calculate (and dynamically recalculate) scene membership. The application’s messages carry with them the metric, path cost function, and threshold, which are sufficient for each node to independently determine whether it is a member of the scene. Each node along a network path determines whether it lies within the scene, and if so, forwards the message. It is possible for a node to qualify to be within the scene based on
multiple paths. These network paths correspond to branches of a routing tree that is set up as part of the distributed calculation of the scene. When a certain data source needs to relay a reply back to the user, the reverse of the path on the routing tree the message took to get to that node can be used. If a node receives a scene message that it has already processed, and the new metric value is not shorter, the new message is dropped. If the new message carries a shorter metric, then the node forwards the information again because it may enable new nodes to be included in the scene.

Scene construction can be formalized in the following way:

*Given a client node $\alpha$, a metric $M$, and a positive threshold $T$, find the set of all hosts $S_\alpha$ such that all hosts in $S_\alpha$ are reachable from $\alpha$ and, for all hosts $\beta$ in $S_\alpha$, the cost of applying the metric on some path from $\alpha$ to $\beta$ is less than $T$. Specifically:*

$$S_\alpha = \langle \text{set } \beta : M(\alpha, \beta) < T :: \beta \rangle$$

In the three-part notation: $= \langle \text{op quantified_variables : range :: expression} \rangle$, the variables from quantified_variables take on all possible values permitted by range. Each instantiation of the variables is substituted in expression, producing a multiset of values to which op is applied, yielding the value of the three-part expression. If no instantiation of the variables satisfies range, then the value of the three-part expression is the identity element for op, e.g., *true* if op is $\forall$, or $\emptyset$ if op is set.

The scene concept conveys a notion of locality, and each application decides how “local” its interactions need to be. A first responder team leader coordinating the team spread throughout the site may want to have an aggregate view of the smoke conditions over the entire site. On the other hand, a particular responder may want a scene that contains readings only from nearby sensors or sensors within his path of movement. The scene for the leader would be “all smoke sensors within the site boundaries,” while the scene for the responder might be “all smoke sensors within 5m.” As a responder moves through the site, the scene specification stays the same, but the data sources belonging to the scene may change.

In the ubiquitous computing scenarios we target, embedded nodes can come and go either because new devices are introduced to the environment, existing devices stopped functioning, or devices moved. Although existing systems treat such situations as exceptional conditions, they must be treated as commonplace. For this reason, the scene abstraction is specifically designed to handle such dynamics. Using the distributed
computation detailed above, the scene implementation automatically recalculates and reconstructs an application’s scene in response to such changes. This adjustment occurs automatically and seamlessly using the application’s provided scene specification (which is now imprinted in the sensor network) and does not involve the application or user directly. Instead, the scene (or the set of devices with which the application interacts) automatically adjusts to reflect the application’s changing operational environment.

To maintain the scene for continuous queries, each member sends periodic beacons advertising its current value for the metric. Each node also monitors beacons from its parent in the routing tree, whose identity is provided as previous hop information in the original scene message. If a node has not heard from its parent for three consecutive beacon intervals, it disqualifies itself from the scene. This corresponds to the node falling outside of the span of the scene due to client mobility or other dynamics. In addition, if the client’s motion necessitates a new node to suddenly become a member of the scene, this new node becomes aware of this condition through the beacon it receives from a current scene member.

Table 1 shows examples of how scenes may be specified. These examples include restricting the scene by the maximum number of hops allowed, the minimum allowable battery power on each participating node, or the maximum physical distance. As one example, SCENE_HOP_COUNT effectively assigns a value of one to each network link. Therefore, using the built-in SCENE_SUM path cost function, the application can build a hop count scene that sums the number of hops a message takes and only includes nodes that are within the number of hops as specified by the threshold. The scene can be further restricted using latency as a second constraint.

Table 1. Example scene definitions

<table>
<thead>
<tr>
<th>Metric</th>
<th>Cost Function</th>
<th>Metric Value</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCENE_HOP_COUNT</td>
<td>SCENE_SUM</td>
<td>number of hops traversed</td>
<td>maximum number of hops</td>
</tr>
</tbody>
</table>

**Battery Power Scene**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Cost Function</th>
<th>Metric Value</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCENE_BATTERY_POWER</td>
<td>SCENE_MIN</td>
<td>minimum battery power</td>
<td>minimum allowable battery power</td>
</tr>
</tbody>
</table>

**Distance Scene**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Cost Function</th>
<th>Metric Value</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCENE_DISTANCE</td>
<td>SCENE_DFORMULA</td>
<td>location of source</td>
<td>maximum physical distance</td>
</tr>
</tbody>
</table>
Intelligent Virtual Sensors

Our middleware paradigm also defines a virtual sensors model (Kabadayi, Pridgen, & Julien, 2006). Virtual sensors provide indirect measurements of abstract conditions (that, by themselves are not physically measurable) by combining sensed data from a group of heterogeneous physical sensors. As an example, the intelligent construction site domain that consists of users with mobile devices distributed over the site and sensors embedded in equipment presents substantial challenges to pervasive computing. For example, users may desire safe load indicators on cranes that determine if a crane is exceeding its capacity. Such a virtual sensor (as shown in Figure 3) would take measurements from physical sensors that monitor boom angle, load, telescoping length, two-block conditions, wind speed, etc. (Neitzel, Seixas, & Ren, 2001). Signals from these individual sensors are used in calculations within the virtual sensor to determine if the crane has exceeded its safe working load. With the virtual sensors model, an application interacts with a combination of physical and virtual devices embedded in the environment.

The physical sensors are the components of the model that provide the physical data types required to compute the desired abstract measurement. Creating a virtual sensor requires defining a group of physical sensors that have some locality relationship (i.e., belong to a local region, where “local” is defined by some property of the network or environment). The resulting virtual sensor has an interface similar to that of a physical sensor (from the application’s perspective).

The virtual sensor hides the explicit data sources from the application, making them appear as one data source that provides the same type of interface as a physical sensor. Our approach to creating a virtual sensor’s declarative specification assumes applications and sensors share knowledge of a naming scheme for the low-level data types the sensor nodes can provide (e.g., “location,” “temperature,” etc.). The types of sensors deployed in a network determine the

Figure 3. An example of a virtual sensor for safe working load
available data types. The programmer, then, only needs to specify the following four parameters for the virtual sensor:

- **Input data types**: Physical (low-level) data types required to compute the desired abstract measurement. Each input data type also includes the number of different sensors (which could be one, two, all, etc.) the virtual sensor would like to obtain this data type from. For example, in deciding if a slab of concrete on a construction site is ready for use, a virtual sensor could combine measurements from concrete heat sensors and strain sensors. The specification of the number of different sensors of a certain type allows us to differentiate between requests for “all of the concrete heat sensors” and “one concrete heat sensor.” The “all” count allows the virtual sensor’s heterogeneous aggregation (over heat and strain sensors) to be combined with homogeneous aggregation (over heat sensors).

- **Aggregator**: A generic function defined to operate over the specific (possibly heterogeneous) input data to calculate the desired measurement.

- **Resulting data type**: The abstract measurement type that is a result of the aggregation.

- **Aggregation frequency**: The frequency with which this aggregation should be made. This frequency determines how consistent the aggregated value is with actual conditions (i.e., more frequently updated aggregations reflect the environment more accurately but generate more communication overhead.). This is similar to the sample frequency of a physical sensor. At each aggregation frequency, the virtual sensor “samples” each physical sensor that it uses and aggregates these results.

By providing these virtual sensor specifications, an application delegates physical sensor discovery to the virtual sensor (and to the framework that supports the virtual sensor). Therefore, if the data sources that support the virtual sensor change over time, the virtual sensor adapts, but the application is not disrupted. This concept of data sources changing over time will be discussed in more detail in the following subsections.

In our model, the input data types carry simply the nature of the physical measure (e.g., “temperature”) provided by a sensor. There could be a data type that is sometimes provided by a physical sensor and is sometimes provided by a virtual sensor (e.g., location provided by GPS (physical sensor) or by triangulation (virtual sensor)). From the application’s perspective, this is only a single data type, and our model uses a data ontology (that may be application domain-specific) to describe these data types. The ontology is basically a simple listing of types. The application programmer can also insert new data types into the ontology to update or augment it over time. Before accessing sensors in the immersive network, the application checks the ontology to determine first if the type can be provided by a physical sensor and, if not, what virtual sensor needs to be deployed to provide the type.

As stated in the introduction, our goal is to support locality of interactions and access to local data. Thus, it is necessary to define the region from which it is allowable to select physical sensors to support a virtual sensor. This allows us to limit the reach of a virtual sensor to some small portion of the network in the immediate vicinity of the client device. When a virtual sensor construction is requested, the necessary scene is created first, and then the virtual sensor is constructed.

After specifying the constraints that build the scene (which consist of a metric definition and a maximum permissible cost for that metric), a client application would like to construct the
virtual sensor using physical sensors from within the scene. That is:

Given the set of hosts \( S_\alpha \) in the scene, the required physical data types \( D_1, D_2, ..., D_n \), the aggregation function \( F \), and the resulting data type, \( D_{\text{res}} \), the virtual sensor can be formalized as:

\[
D_{\text{res}} = F(D_1, D_2, ..., D_n) \text{ where } \langle \forall D_i : 1 \leq i \leq n : (\exists S : |S| = D_i.\text{count} \land (\forall s \in S : s \in S_\alpha \land D_i.\text{type} (\triangleright s)) \rangle
\]

The “\( D_i.\text{type} (\triangleright s) \)” construct denotes the fact that “sensor \( s \) can provide the data type specified in \( D_i.\text{type} \).” In the above definition, the set \( S \) is the subset of \( S_\alpha \) that defines which physical devices contribute to the virtual sensor. If the construct in the last line of the definition evaluates to false, it is not possible to construct the specified virtual sensor.

As stated in the scene definition, the input data types (\( D_1, D_2, ..., D_n \)) are defined by the type of data they request and the number of independent readings of that type that are required. We assume the former is expressed as \( D_i.\text{type} \) and the latter is expressed as \( D_i.\text{count} \). For example, the virtual sensor shown in Figure 3 requires two data types and one sensor of each type: a single crane base position and a single crane boom position. On the other hand, a virtual sensor that generates the average temperature of a curing pad of concrete requires temperature values from \( n \) temperature sensors. In this case, only one \( D \) is provided, and its count value reflects the number of sensors to be polled. As we described previously, the count value included in the declaration of the virtual sensor can be a number (e.g., one as in the case of the crane sensor above) or \( \text{all} \), indicating that all matching sensors in the scene should be polled. In the latter case, the count value for the data type is set to be exactly the cardinality of the scene, \( |S_\alpha| \).

To summarize, the virtual sensor aggregation function \( (F) \) operates on the input data that is of the specified type to yield the specified output data type. This function evaluation takes place at every interval specified by the aggregation frequency.

Dynamic sensor discovery, that is the discovery of the virtual sensor by the application, takes place on the basis of the virtual sensor specification. Virtual sensors provide the same interface to the user as physical sensors. The data type ontology that was mentioned above exists at this interface and defines the data types available to the application. A domain expert can create complex virtual sensors by hand and add their types to the ontology. Most importantly, the application does not have to know that it is discovering a virtual sensor instead of a physical sensor. The developer selects a data type listed in the ontology from a scene. If a physical sensor can provide that data type, no virtual sensor construction is necessary. Otherwise, the virtual sensor is activated and searches for supporting physical sensors in the scene. The ontology provides this mapping so that the application developer using virtual sensors must do nothing more than select the virtual sensor associated with the type in the ontology and request its deployment. This is handled automatically by the middleware at the time an application requests data of a particular type.

If a virtual sensor is being used to obtain periodic responses, it needs to be dynamically refreshed. At every refresh interval specified by the virtual sensor’s aggregation frequency, the virtual sensor gets new measurements from each physical sensor contributing to it and recalculates the virtual measurement. During its lifetime, some sensors contributing to a virtual sensor may deplete their battery power and become nonfunctional. In this case, the virtual sensor attempts to discover a new sensor that can provide the data that sensor was providing; if another such sensor does not exist in the scene, the virtual sensor fails.

Furthermore, while the sensor nodes used in pervasive environments are embedded (hence stationary), the application interacting with them runs on a device carried by a mobile user.
Therefore, the device’s connections to particular sensors and the area from which the application desires to draw information (i.e., the scene) are subject to constant change. These changes need to be seamlessly handled without revealing the underlying dynamics to the user. The dynamics associated with user movement may cause the physical sensors that comprise the virtual sensor to change locations or become unavailable.

Figure 4. The dynamics associated with user movement

Figure 5. Abstract depiction of a virtual sensor that uses n physical nodes: (a) n physical sensors, aggregated into a virtual sensor which runs locally, (b) n physical sensors, aggregated into a virtual sensor which runs remotely
Enabling Programmable Ubiquitous Computing Environments

to change and are handled by the scene abstraction (see Figure 4). Since the physical data sources that can contribute to a virtual sensor are restricted by the scene, any data source that satisfies the data requirements from that scene is a valid data source for the virtual sensor.

To summarize, this dynamic maintenance allows the user to interact directly with a changing set of local information sources. The virtual sensor hides the underlying complexity from the user.

Figure 5 abstractly depicts n physical sensors, aggregated into a virtual sensor which can run (a) locally (on the client) or (b) remotely (in the network). The physical sensors are illustrated using circles, and the different shadings indicate their heterogeneous nature. The sensors inside the large dashed circle contribute to the virtual sensor, but only a few have been shown with arrows representing the data they send back, for ease of presentation. The virtual sensor code can either run locally on the client device or be deployed to a resource-constrained sensor within the network. When deployed remotely, this code will be dynamically received by a listener on the remote sensor and executed. If all of the physical sensors are in a cluster, and that cluster is several hops away from the user’s device, then it may make sense to send the virtual sensor out to the cluster. On the other hand, if each of the sensors that make up the virtual sensor is within one hop of the user, then the virtual sensor should run on the user’s device. With respect to the application interface, it does not matter if the virtual sensor is deployed on the client’s device or remotely in the network, but it might improve performance to use a certain option depending on the application’s situation.

Benefiting from Middleware in Ubiquitous Computing

In this section, we present a complete middleware solution, the DAIS middleware model, which incorporates the scene and virtual sensor abstractions, adheres to the requirements enumerated in the previous section, and provides a cohesive environment for ubiquitous computing application development. Figure 6 shows the DAIS architecture, which consists of a handheld component (running on, for example, a laptop or a PDA) and the immersive sensor environment (defined by a community of sensors). The figure shows DAIS’s explicit hierarchical model which enables more powerful devices (i.e., client devices) to support more of the middleware’s functionality than resource-constrained devices (e.g., physical sensors). As Figure 6 shows, a client application runs with the support of the two key abstractions introduced in the previous sections, namely scenes and virtual sensors. The types of queries possible in DAIS can be classified into one-time queries (which return a single result from each participating sensor) and persistent queries (which return periodic responses from participating sensors). To support these types, we provide two different methods for posing queries to the network. We also include versions of these two methods that request processing of the retrieved data before the result is handed back to the application.

In the remainder of this section, we describe the use of the strategy pattern to perform communication protocol encapsulation, how queries come from the application, and what happens as these queries travel through the middleware depicted in Figure 6. Figure 7 shows a simplified object diagram of the DAIS middleware layers. The names of the layers on the right of the figure correspond to the layers in Figure 6.

Strategy Pattern Interface

As shown in Figure 7, our middleware makes use of the strategy pattern (Gamma, Helm, Johnson, & Vlissides, 1995), a software design pattern in which algorithms (such as strategies for query dissemination) can be chosen at runtime depending on system conditions. The strategy pattern provides a means to define a family of algorithms,
Figure 6. The DAIS high-level architecture. The left-hand side shows the components comprising the model on the component carried by the user (e.g., PDA or laptop), and the right-hand side shows the DAIS components on the sensors.

Figure 7. DAIS object diagram
Encapsulate each one as an object, and make them interchangeable. Such an approach allows the algorithms to vary independently from clients that use them.

In DAIS, the clients that employ the strategies are the queries, and the different strategies are the SceneStrategy algorithms. These algorithms determine how a query is disseminated to the scene and how the QueryResult is returned. If a particular dissemination algorithm other than the default is required for a specific application, an appropriate SceneStrategy algorithm is instantiated.

Two principle directives of object-oriented design are used in the strategy pattern: encapsulate the concept that varies, and program to an interface, not to an implementation. Using the strategy pattern, we decouple the Query from the code that runs it so we can vary the query dissemination algorithm without modifying the Query class. The loose coupling that the strategy pattern enables between the components makes the system easier to maintain, extend, and reuse.

For now, we provide only a single implementation of the strategy, the BasicScene. This is a simple, greedy scheme in which all data aggregation is performed locally. We have chosen this as a first step to provide a quick prototype of the entire middleware. Other communication approaches can be swapped in for the BasicScene (for example, one built around TinyDB (Madden et al., 2005) or directed diffusion (Intanagonwiwat, Govindan, Estrin, Heideman, & Silva, 2003), although the implementations of these approaches on the sensors may have to be modified slightly to accommodate scene construction). By defining the SceneStrategy interface, we enable developers who are experts in existing communication approaches to create simple plug-ins that use different query dissemination and/or aggregation protocols. Different communication paradigms can be used in different environments or to support different application domains depending on the resource constraints or domain-specific capabilities of the devices in a particular domain.

Each SceneStrategy interacts with the javax.comm package to provide the DAIS abstraction protocols that allow the portion of the middleware implemented in Java (described above) to interact with the sensor hardware. Each SceneStrategy requires not only a high-level portion implemented on the handheld device, but also a low-level portion that runs on the sensors. Next, we briefly discuss our BasicScene implementation that runs on the sensor nodes that respond to a client application’s queries. Complete details of this implementation can be found in (Kabadayi & Julien, 2007). This serves as just one example of a particular implementation of the SceneStrategy.

In DAIS, we have developed the sensor components for the Crossbow MICA2 mote platform (Crossbow, 2007). Our initial implementation is written for TinyOS (Hill et al., 2000) in the nesC language (Gay et al., 2003) and helps the BasicScene strategy conform to the SceneStrategy interface.

Building Scenes

In nesC, an application consists of modules wired together via shared interfaces to form configurations. We have created several components that form the fundamental functionality of the Scene configuration. Figure 8 abstractly depicts the necessary components and the interfaces they share.

This implementation functions as a routing component on each node, receiving each incoming message from the radio and processing it as our protocol dictates. In this picture, we show components as rounded rectangles and interfaces as arrows connecting components. A component provides an interface if the corresponding arrow points towards it and uses an interface if the arrow points away from it. If a component provides an
interface, it must implement all of the commands in the interface, and if a component uses an interface, it can call any commands in the interface and must handle all events generated by the interface. In short, the SceneM component uses the messages sent as part of the scene construction protocol to determine whether the local node is contained within the specified scene. If so, the application-level message is propagated to the application layer on the sensor, and the scene construction message is propagated to the neighbors to allow includes of additional nodes.

Determining whether or not it is part of a scene may require the SceneM component to access local context information (through the ContextSources in Figure 8). For example, if a scene is defined based on physical locations, each node must access a location context sensor to determine the node’s physical location.

When a node is part of a scene over which persistent queries have been issued, a Beacon message proactively maintains the connectivity among the scene members. If, in the exchange of these Beacon messages and monitoring the ContextSources that are in use, a node discovers that it no longer qualifies to be a member of the scene (e.g., the scene is based on physical location and the client device has moved), the node removes itself from the scene and ceases this proactive behavior.

Figure 8. Implementation of Scene functionality on sensors
Enabling Programmable Ubiquitous Computing Environments

Processing Dynamic Queries:
A Step-By-Step Example

To describe the DAIS model and how it works, we follow a query from the application developer’s hands all the way into the network to a participating sensor and back. To present the steps involved in the process, we use a single, specific application example taken from the intelligent construction site domain. In the example we have selected, the application’s user would like to receive a notification if any crane load reading within 100m exceeds a safe threshold. The user would like this information to be updated every 10 seconds to ensure he has the most current information.

• **Step 1: Create a Virtual Sensor.** As depicted in Figure 3, whether or not a crane load is safe depends on several conditions and is not simply a factor of the weight of a load. In a simple scenario, we assume that the safety of a load depends on the position of the load on the crane arm (e.g., heavier loads should be kept closer to the center of the crane), the tension on the boom (e.g., heavier loads and windier conditions both imply greater tensions, which can increase danger), and the acceleration of the crane (e.g., moving heavier loads quickly is dangerous). An application domain expert (with knowledge about crane safety) can create a virtual sensor that collects these aspects of the crane environment and generates a crane safety value (“safe” or “unsafe”). The following code would define the virtual sensor:

```java
VirtualSensor craneVS = new VirtualSensor({new CraneArmTension(),
    new CraneBoomPosition(),
    new CraneAcceleration()},
    new LoadAggregator(),
    new CraneSafety(),
    1);
```

Within this definition, the first argument to the constructor contains an array of the virtual sensor’s input data types. We omit the count for each type for brevity; in this example, the virtual sensor requires only one data reading of each type. The second argument contains a function definition that dictates how these values should be combined to generate the crane safety type. The third argument specifies the return type of the virtual sensor; in this case, whether or not the crane is safe. The final argument, an integer, gives the desired aggregation frequency for this virtual sensor. In this example, the physical sensors supporting the virtual sensor should send updates of their values every second. The application domain expert must define the code for the `LoadAggregator` and the meaning of the `CraneSafety` type; the other types are assumed to be physical measures that can be provided within the deployed network. The definition of this virtual sensor must also be added to the data ontology so applications needing to query the network can locate its type and automatically deploy it when it is needed. This is accomplished through the following code, which associates the `CraneSafety` type with the `craneVS` virtual sensor definition.

```java
ontology.add({new CraneSafety(), craneVS});
```

This allows applications searching the data ontology to locate the virtual sensor functionality when they request the `CraneSafety` type.

• **Step 2: Declare a Scene.** This is the first step is performed by an application developer when creating and interacting with an immersive sensor network through the programming interface. Nothing happens involving network communication until the application actually uses the scene,
enabling programmable ubiquitous computing environments

Reducing communication overhead. For our application example, the developer uses the following code to invoke the constructor of a Scene object that defines a scene that includes every sensor (not just those measuring crane load safety) within 100m of the declaring device:

```java
Scene s = new Scene(new Constraint(Scene.SCENE_DISTANCE, Scene.SCENE_DFORMULA, new IntegerThreshold(100)));
```

When the application subsequently needs to query the constructed scene, it calls the `getSceneView()` method, which returns a handle to a SceneView. The application can then use this SceneView to send a query over the scene.

- **Step 3: Issue a Query.** This step is performed by the application developer using the SceneView instance created and accessed in the previous step. In our example application, the developer must first create the query:

```java
Query q = new Query(new Constraint("CraneSafety", Query.EQUALS_OPERATOR, "unsafe"));
```

In this case, the Query is defined by a single Constraint that requires returns any crane safety value of “unsafe” within 100m to be returned to the developer. After creating the Query above and a ResultListener r to receive the results (omitted for brevity), the application developer dispatches it using the SceneView:

```java
SceneView sv = s.getSceneView();
int receipt = sv.registerQuery(q, r, 10);
```

where 10 refers to the fact that the application wishes to sense the safety of crane loads every 10 seconds. Upon receiving any query request, the SceneView object adjusts its state in several ways. First, for every query, a table within the SceneView is updated with a mapping from a unique query id generated for the query to the ResultListener handle provided with the query. In addition, for persistent queries, this unique id is returned as a receipt of registration that can be used in subsequent interactions to, for example, deregister the query.

- **Step 4: Initialize Necessary Virtual Sensors.** Before dispatching the query to the scene, the middleware checks the data ontology to determine if the request type (in this case CraneSafety) is a primitive type in this network (i.e., it can be provided by a physical sensor), or if it is a virtual type for which a virtual sensor must be deployed. In the latter case (and in the case of this example), it is at this instant that the middleware creates the virtual sensor in the network. It is possible that multiple virtual sensors could be supported within the scene (e.g., in our scenario, it is possible that there are multiple cranes within 100m). For now, our implementation assumes that only a single virtual sensor is constructed in the scene; the final section of this chapter discusses the potential research necessary in this area in more detail. In our current implementation, the virtual sensor is “deployed” locally (as in Figure 5(a)), and this step connects the remote physical sensors to the local virtual sensor. The virtual sensor is now present within the scene and can be discovered by subsequent queries (from this node or other nodes). It is available to the QueryProcessor as one of the Sensors shown in Figure 9 and described below.
• **Step 3: Create Local Data Proxy.** Control has passed from the application developer to the middleware, which is now responsible for ensuring that the application’s `ResultListener` is called with the appropriate data at the appropriate times. DAIS creates a local proxy dedicated to handling return calls for this query. The local data proxy is especially important in facilitating the translation between the low-level language spoken by the sensors (nesC in our implementation) in the network and the high-level language the application uses (Java).

• **Step 4: Construct and Distribute Protocol Query.** The local data proxy with the DAIS middleware transforms the application’s request into a protocol data unit for the scene implementation in use. As was shown in Figure 7, several different protocols can provide the communication functionality as long as they adhere to the specified strategy pattern interface. The scene implementations must handle both persistent and one-time queries, as dictated by the strategy pattern interface. In our current implementation of the scene protocol (Kabadayi & Julien, 2007), the scene protocol message carries the information about scene membership constraints and the data query at the same time. This reduces the communication overhead by constructing the scene on demand. The details of the communication protocol are omitted here (see the referenced paper); it suffices to say that, by its definition, the protocol ensures that the data query is delivered to exactly the set of sensor nodes that satisfy the scene’s constraints. In our example application, this
Enabling Programmable Ubiquitous Computing Environments

means that every sensor within 100m will receive the data query constructed above. Every sensor receiving this query (i.e., all sensors within 100m) that supports the requested type (i.e., that have a CraneSafety sensor) periodically evaluates the query, but responds only if and when the sensor returns exactly the value unsafe.

- **Step 5: Process Scene Query on Remote Sensor.** We use TinyOS to implement functionality on the sensors; when the communication protocol receives and processes a scene message, if it determines that the node is within the scene, it passes the received message up to the application layer. In DAIS, the QueryProcessor, implemented in TinyOS, defines this application layer. This is essentially a slimmed down version of our middleware that is capable of running on the resource-constrained device. An abstract representation of the TinyOS implementation of the QueryProcessor is shown in Figure 9, which uses the same notational conventions used in Figure 8.

In our implementation, the QueryProcessor component provides the functionality shown at the top layer of the sensor portion of the architecture in Figure 6. The query arrives in the QueryProcessor through the receive event of the Receive interface. If the query is a one-time query (as indicated by a field in the TinyOS message), then the QueryProcessor simply connects to the onboard sensor that can provide the requested data type (depicted as Sensor in the figure). If the data request is for a sensor type that is not supported on this device (i.e., the sensor table stored in the QueryProcessor has no mapping to a sensor that can provide the specified data type), then the message is ignored. This sensor table includes both physical and virtual sensors available at this node. Even if the node ignores the application layer message because it cannot satisfy the data query, the node is still included in the scene because it may be a necessary routing node connecting the requester to another node that does have the required sensor.

If the query is persistent (as in our safe crane load example), then in addition to immediately returning the requested value if the query is satisfied, the QueryProcessorM module also initializes a QueryTimer using the request frequency specified in the data portion of the received message. When the timer fires, QueryProcessorM retrieves a value from the sensor and sends it back to the initial requester using the SendMsg interface of the QueuedSend module.

When a node is no longer in a scene (either because it moved, the client moved, or something happened in between to cause the scene membership requirement to no longer be satisfied), the scene communication implementation creates a null message that it sends to the QueryProcessor through the Receive interface. The QueryProcessor takes this message as a sign to cease streaming data back to the requester and stops the QueryTimer.

- **Step 6: Reply from Query Processor.** If the query processor possessed the correct sensor and the other components of the query are also satisfied (e.g., the constraints on the data value), the query processor replies (and continues to reply periodically to a persistent query). In our specific application scenario, every sensor node in the scene (i.e., within 100m) receives the Query for the CraneSafety type. In this case, only one of the sensors, the one supporting the virtual sensor, can actually provide the type. This QueryProcessor on this sensor monitors the values the virtual sensor generates over time (every 10 seconds) and returns a message to the application if the
CraneSafety value is ever “unsafe.” If the value is “safe,” the node does not reply to the client application. Replies are sent through the SendMsg interface shown in Figure 9 and use basic multihop routing to return to the original requester.

- **Step 7: Distribute Result to Client Device.** After propagating through the underlying communication substrate, query replies will arrive at the client device’s sensor network interface. At the client device, the result is demultiplexed by the Request Processor (shown in Figure 6) and handed off to the appropriate local proxy. Again, the local proxy is automatically generated and managed by the middleware. It translates the low-level query reply into a high-level Result and invokes the application’s registered resultReceived() method. It is important to note that, as shown in Figure 7, multiple queries may be active over a single scene at any given time. For each scene, the SceneView controls all of these queries and connects them to the underlying implementation via the strategy pattern interface. When the result is passed off to the ResultListener, control transfers back to the application which handles the query’s results (or queries’ results if multiple matches existed). For persistent queries, as more results arrive, the same process occurs until the application deregisters the query.

This complete example demonstrates the few steps the application developer must accomplish to be able to leverage the automatic capabilities of the DAIS middleware. By simply interacting with the straightforward APIs the middleware provides, the application developer can define virtual sensors, create scenes over which to interact with a network, and dispatch queries. The persistent query capabilities of the DAIS middleware enable the application to monitor potentially dynamic situations as the user moves through an immersive sensor network. Both the scene and virtual sensor abstractions are capable of automatically adjusting to such changes and making the application’s perspective on the world reflect the instantaneous operating conditions.

By coupling communication constructs (through the scene communication protocol) in addition to programming abstractions (through the virtual sensors and scene programming interfaces), the DAIS middleware exposes significant expressiveness to the application developer while accepting the burden of implementing the complex interactions in the ubiquitous computing network. From the application developer’s perspective, the most complicated portions of interacting with the DAIS middleware lie in properly defining the virtual sensors and creating meaningful queries to issue over scenes. Both of these aspects require significant domain-specific knowledge, and the simple APIs we have designed allow domain experts to easily encode this knowledge which the DAIS middleware can subsequently translate to low-level code to run in the sensor network.

**CONCLUSION**

Given the extreme demands for simplified programming of ubiquitous computing applications, it is obvious that programming environments tailored to this domain are required. In this chapter, we have first argued that immersive sensor networks are essential to the future success of ubiquitous computing. We have then surveyed existing work in the area of providing simplified programming constructs for application developers working in both sensor networks and ubiquitous computing. Finally, the chapter has presented a new paradigm for immersive sensor networks that combines traditional sensor network technologies with the vision of ubiquitous computing.

Because any middleware for immersive environments must operate at least to some degree on resource-constrained devices, it is important to achieve reasonable performance with respect to
battery lifetimes and communication overhead. Our work on the scene protocol (Kabadayi & Julien, 2007) has shown that, by localizing communication to a region surrounding the application in question (as depicted in Figure 1(b)), we greatly reduce the energy and communication impact on other devices in the network. In addition, this approach provides a localized view of the world, which is invaluable to ubiquitous computing applications. A similar performance characterization of the query processing capabilities on the sensor nodes themselves is underway; here we must ensure that responses (especially those indicating anomalous conditions) can be returned to client applications in a timely fashion.

Finally, the example application given in the previous section demonstrates the ease with which application developers can interact with the DAIS middleware constructs. By employing a few straightforward middleware interfaces, applications can easily define regions of interactions that expand and contract in response to context properties of the environment. Applications can also define expressive queries that the middleware dispatches to the network, allowing in-network query resolution, which can reduce communication overhead and latency.

Given the systems available as the current state of the art, we are lacking the ability to quickly and easily design, develop, and deploy expressive applications that allow localized interaction and coordination among users immersed in ubiquitous computing environments and the environments themselves. The DAIS middleware, presented in this chapter, takes a first step towards providing such an environment that focuses on localized interactions that are based on relative notions of locations between devices.

**FUTURE DIRECTIONS**

As our survey of existing research has indicated, there is a significant lack of availability of tools and development environments that enable application development in ubiquitous computing environments. Clearly, significant progress must be made in this area before applications can be rapidly and reliably created for ubiquitous computing environments. Such tools should be motivated directly by the characteristics specific to ubiquitous computing environments as detailed in this chapter, and they must provide support for coordination among users immersed in the environment. Such tools must allow applications to remain highly expressive while embedding much of the complexity within a middleware that eases the development burden.

Evaluation of the results of this research must also come to the forefront. Testing and evaluation of development constructs is inherently difficult; ubiquitous computing environments add extra concerns to the mix. New mechanisms for supporting easy and correct evaluation and testing must be developed that are tailored to the ubiquitous computing environment. Specifically, in addressing the above development environments, we must be able to evaluate the quality of the abstractions with respect to aiding application development.

Finally, in evaluating the coordination requirements for ubiquitous computing environments, an obvious central theme is the ability to group embedded nodes based on relative physical and logical properties. The scene abstraction is one example of this, and it is a natural approach to defining regions of coordination around an immersed application user. We have also used this scene approach to support the virtual sensor abstraction, but it is not as natural in this case. To enable abstractions such as the virtual sensor to accurately reflect an environment, new neighborhood abstractions must be developed that enable sensor nodes to self-organize in response to application queries.

In conclusion, while many positive steps have been made in enabling technologies for immersive sensor network support of ubiquitous computing
Applications, several research aspects remain open and must be addressed before we will see widespread adoption of ubiquitous computing networks and applications.

ACKNOWLEDGMENT

The authors would like to thank the Center for Excellence in Distributed Global Environments for providing research facilities and the collaborative environment. This research was funded, in part, by the National Science Foundation (NSF), Grants #CNS-0620245 and OCI-0636299. The conclusions herein are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

REFERENCES


Enabling Programmable Ubiquitous Computing Environments


Index

A

ABCNews.com 826
abstract syntax 885–886
academic advising problem 1121
access points (APs) 264
access to technology 1510
access-control engine 275
accessibility 194, 1632
accessibility, definition of 226
ACQUA project 794
ACQUA, evaluation of 800
action direction 2082
action modus 2083
action objects 2075
action science 1829–1832
action workflow (AW) 2076
actions 2393
actions in contract phase 2084
actions, automatic 2394
activation activity 953
active learning model 1662
activity analysis 678
activity checklist 916
activity diagrams 407–409
activity diagrams conditions 409
activity diagrams description 408
activity diagrams transitions 408
activity theory 583, 913
activity theory framework 913
activity theory model 932, 933, 934, 936
activity-oriented design method (AODM) 518
actor network theory (ANT) 583, 2399
actors 2392
ACT-R cognitive architecture 1156
adaptability 1507, 1630, 2542, 2545
adaptation 393
adaptive content 1631
adaptive educational environments 1388
adaptive features 797
adaptive methods 169
adaptive multimodal architectures 1009
adaptive personalization systems 176
adaptive site behavior 181
adaptive system 395, 792
adaptive Web site 791
adaptivity 1630
Adobe Dreamweaver 2605
Adobe Indesign 2605
Adobe Photoshop 2605
adoption process 1061
advanced methods of software facilitated communication, definition of 1870
advising procedure 1126
aesthetic criteria 378
aesthetics 383
affective responses to attitude 2288, 2294
AFrame 714
age of mobile communication, being a child 2655
age of mobile communication, being a parent 2655
agent technology 1070
agents 378, 2395
aggregate usage profiles 738
AHA project 1634
Amazon.com 170
amberfication paradox 2107
amberfication paradox, definition of 2116
ambient business 2191, 2193, 2195, 2196, 2197, 2199, 2201, 2206, 2207
ambient business settings, open innovation in 2199
ambient business, drivers for open innovation in 2198
ambient business, emerging technologies 2194
ambient business, framework for 2193
ambient business, open innovation 2201
ambient devices 271
ambient intelligence (AmI) 1, 21, 328, 2192
ambient intelligence (AmI), definition of 2626
ambient media 22
American health care system 1688
Americans with Disabilities Act (ADA) 2543, 2544, 2555
Americans with Disabilities Act (ADA), definition of 2555
anachronism 2332
analytic hierarchy process (AHP) 1858
androcentric dominance online 2280
angle of arrival (AoA) 1175, 1178, 1180, 1181, 1182, 1187, 1188, 1189, 1190, 1191, 1192, 1194, 1195, 1197
animal disease 1658
Anoto Pen 271
anthropology 133, 631
anti-social attacks 2678
AOL Instant Messenger (AIM) 2612
Apple.com 826
application areas 262
application composition layers, business 277
application scenarios 271
application service providers (ASP) 1383
applied IT 1743
applied technology perspective 981
approaches, activity-centered 22
AR-arrangement 633
architectural personalization 853
architectures, feature-level 97
architectures, peer-to-peer, definition of 261
architectures, plug-in-based 239
architectures, software 97
architectures, software, definition of 227
architectures, software, for content separation 222
artifacts 2392
artifacts, definition of 314
artificial intelligence (AI) 2, 4, 6, 7, 8
assistive technologies 2545, 2552
association rules 799, 2171
association rules semantics 888–889
association rules syntax 888–889
A-structure, theory of 2502
asynchronous communication 1865
asynchronous communication, definition of 1870
asynchronous online communication, definition of 298
asynchronous transfer mode (ATM) 268
attention investment model 28
attentional capacity, definition of 959
attentive user interfaces (AUIs) 630
attitudes 66, 2288, 2294
audience reception workshop 443
augmented reality 629, 636
authentication 326
authentication, private 347
authoring 631
authoring tools 630, 1368
autocasting 2540
autocasting, definition of 2540
automated speech recognition (ASR) 1216
automated teller machines (ATMs) 109
autonomy 2659, 2665, 2666, 2667, 2668, 2670
auto-personalization Web pages 807
Autostadt project 2466
Autostadt, IT and postmodernity 2468
AVANTGARDE 639, 642
avatar, definition of 298
avatars 298
average variances extracted (AVE) 862
Avex Records 1224

B
baby boomers 1687
balli, balli 1204
bar-coded ticket 1639
Barger, J. 1292
baseline behavior 1974
basic methods of computer-mediated communication, definition of 1870
BDI architecture, definition of 2626
behavioral patterns 136
behavioral task analysis 905
bibliographic coupling, definition of 314
bibliometrics, definition of 314
binary-phase-shift keying (BPSK) 1039
blended learning, definition of 322
blocker tags 354
blogcensus 1488
blogcount 1488
bloggers 1485
blogosphere 1487
blogroll, definition of 1297
blogrolling 1293
blogs 1251, 1252, 1255, 1257, 1258, 1259, 1261, 1307, 1313, 1408, 1484, 1485, 2181, 2594, 2610
blogs, academic 1291, 1292, 1293
blogs, communication in 2183
blogs, communication, nature of  2183
blogs, community  1489
blogs, controversies  1255, 1255–1257
blogs, definition of  1297, 1307, 2189, 2540
blogs, education and research  1489
blogs, education portals  1298
blogs, elementary and middle school  1493
blogs, features of  1486
blogs, high school  1492
blogs, in a multinational firm  2185
blogs, issues in 1255
blogs, journalism  1489
blogs, knowledge  2182
blogs, knowledge in 1490
blogs, knowledge, definition of  2189
blogs, knowledge, firm internal use of  2181
blogs, organizational culture of  2184
blogs, organizational, definition of  2189
blogs, personal  1489
blogs, problems associated with  1255
blogs, role of technology in  2184
blogs, trust in  2185
Bluetooth  1038, 1089
blurring data  346
Boingo  264
boomers and seniors, health of  1690
Bradford partitions, definition of  314
Brazilian System for the Digital Television (SBTVD)  449
British Council, facilitating knowledge sharing  1905
British Telecommunications  268
browser-based architecture  1216
burst detection, definition of  1685
business action  2075
business action theory (BAT)  2071
business action theory (BAT) model  2072, 2074
business action theory (BAT) model, assessment phase  2085
business action theory (BAT) model, contractual phase  2084
business action theory (BAT) model, essential concepts  2075
business action theory (BAT) model, fulfilment phase  2084
business action theory (BAT) model, overview  2074
business context  1906
business interaction, levels of  2074
business modeling language  2399
business phase  2076
business process reengineering (BPR)  2142
business reengineering  2139
business transaction  2074, 2076
business-to-business (B2B)  2075
business-to-consumer (B2C)  2075

C
CAD drawing  494, 495, 502
calm technology  229
captology  2336
CAR  636, 637, 639, 642
carrier sense (CS) medium access with collision avoidance (CSMA-CA)  1039, 1043
CASE tools  1940
category management  808, 812, 815
ceiling  642
Center for Lifelong Learning and Design (L3D)  704
center for universal design  2109
central controller  274
centralized provision  283
centralized service provision  283
centralized service provision, example of  284
cerebral lateralization  2277
Chaku-melo  1224
Chaku-uta  1224
channel  1823
chat room  1823
chatiquette  320
chatiquette, definition of  322
checklists  457
Chinese collectivism culture  1793, 1795
Chinese cultural values  1786, 1795, 1797
Chinese POS disambiguation  1595
Chi-square test, definition of  1250
choiceboard  1677
choiceboard user satisfaction  1956
choiceboard, definition of  1677
chronemics  1820
ciphers  340
Cité des Sciences et de l’Industrie  1639
CiteSeer  737
class definition of  203
class librarie  631
classical language lab configuration  1365
classification  1083
classification algorithms  2170
classification of degree, definition of  1589
classroom teaching  1436
clickstream  807, 815, 1086
clickstream, definition of  815
click-through  178
clustering  877–879, 881–888, 1083, 2170
clustering algorithms  2170
clustering task 892
clustering, syntax of 889
cluster-oriented genetic algorithm (COGA) 385
CNN.com 826
cocitation analyses, definition of 314
code generation 417–428
code generation GUI components 420–428
code generation mapping states 422–428
code generation mapping states, terminal 424–428
code generation mapping transitions 425–428
code generation non-terminal mapping states 422–424
code generation techniques 405
code generation user-interfaces diagrams 418–428
codified knowledge 2040
coding system, validation of 834
cognition 2546
cognitive abilities 2481
cognitive activities 951, 952
cognitive dissonance 575
cognitive engineering 62, 68, 900
cognitive informatics (CI) 33
cognitive informatics (CI) laws of software 39
cognitive informatics (CI), applications of 47
cognitive informatics (CI), denotational mathematics for 41
cognitive informatics (CI), fundamental theories of 35
cognitive informatics (CI), theoretical framework for 33–59
cognitive needs 465
cognitive perception 66
cognitive processing models 2546, 2547, 2552, 2555
cognitive processing models, definition of 2555
cognitive psychology 2405
cognitive responses to attitude 2288, 2294
cognitive science 466, 904
cognitive spatial abilities 2481
cognitive style 2277
cognitive task analysis (CTA) 203, 488, 905, 908
cognitive task design (CTD) 75
cognitive task models 905, 909
cognitive trapdoor games 326
cognitive work analysis (CWA) 908
Colby, Bates, and Bowdoin (CBB) plagiarism project 1251, 1254, 1258
collaborating partners 2044
collaboration 1491, 2316
collaboration challenges 1929
collaboration in networks 1915
collaboration, passive 335
collaborative design space explorations 636
collaborative filtering (CF) 733, 808, 815, 1630
collaborative filtering (CF), definition of 815
collaborative learning 1336
collaborative product development 2042
collaborative software 1384
collective meaning 1873
collectivism 1571, 1786
collectivist culture 1826
color theory 179
comment, definition of 1297
commercial personalization 854
communication 220, 439, 2395
communication and trust 2126
communication channel protection 2424
communication channels 1952
communication media (CM) 1975, 1994, 2122
communication medium 1819
communication, human-to-human 225
communication, media of 2125
communication, model of 2124
communication, nonverbal 2124
communication, visual 2124
communications 2124
communications, bootstrapping secure 349
communications, human 221
communities of practice (CoPs) 1044, 2045, 2052
community networks, collaboration challenges for 1918
community networks, phases of development 1919
community telecommunication networks, collaboration challenges 1915
complementary modalities 107
comprehensive business reengineering 2139
comprehensive interactive learning environment (CompILE) 698–712
computational resources 280
computer attitude scale 2294
computer interaction 194
computer interface 434
computer kids 324
computer literacy 1818
computer science (CS) 1583, 1584, 1743
computer socialization 2280
computer system design 2675
computer system verification 2438
computer system, building trust in 2440
computer usage 1438
computer user satisfaction inventory (CUSI) 363, 374
computer-assisted learning (CAL) 1656
computer-based classroom teaching 1436
computer-based information technology 2325
computer-human interface design, definition of 2116
computer-mediated communication (CMC) 1334, 1547, 1745, 1752, 1817–1818, 1818, 1837, 1838, 1839, 1840, 1841, 1842, 1843, 1844, 1845, 1846, 1847, 1848, 1849, 1850, 1864
computer-mediated communication (CMC) 2118
computer-mediated communication (CMC) literature on 299, 1550
computer-mediated communication (CMC) research on 299
computer-mediated communication (CMC), approaches to 2182
computer-mediated communication (CMC), characteristics of 317
computer-mediated communication (CMC), common core terms for 306
computer-mediated communication (CMC), conceptual map of research methodologies 307
computer-mediated communication (CMC), definition of 298, 314, 322, 1848, 1870
computer-mediated communication (CMC), effects of 1864
computer-mediated communication (CMC), group decision making 1849
computer-mediated communication (CMC), impression formation cues in 1838
computer-mediated communication (CMC), impression formation in 1837
computer-mediated communication (CMC), key benefits of 1851
computer-mediated communication (CMC), linguistics of 1819
computer-mediated communication (CMC), meaning of 316
computer-mediated communication (CMC), online introduction to 295, 296
computer-mediated communication (CMC), participant orientation 294, 296
computer-mediated communication (CMC), preparing participants for 290
computer-mediated communication (CMC), strategies for managing impressions in 1842
computer-mediated communication (CMC), theories for 318
computer-mediated communication (CMC), unique core terms for 306
computer-mediated communication (CMC), virtual learning communities 316
computer-mediated communications (CMC), trust in 2118
computer-mediated environment 1970
computer-mediated trust 2439
computer-operated voice 168
computer-supported team-based learning 1341
computer-supported team-based learning events 1340
computing context 236–237
computing courses, achievements in 1584
computing, end-user 25, 32
computing, mobile 334
computing, nomadic 1016
computing, pervasive 19
computing, ubiquitous (UC) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18
computing, ubiquitous (UC), synonyms of 7
conative responses to attitude 2288, 2294
concepts, definition of 314
conceptual design workshop 443
conceptual framework 133
conceptual model 2154
conceptual modeling 1102
conceptualization 2148
conceptualizer 2028
confidence 26, 27, 185
confirmatory factor analysis 1961
conformity assessment, definition of 2589
Confucian dynamism 1204
Confucian influence 1795
Confucian teachings 141
Confucian work dynamism 141
connectivist environment, affordances of learning in 963
connectivist learning activities 961
consensual relationships 136
consequential actions 2394
construction and repair agents (CARAs) 381
constructivism 1307, 1467
constructivism and learning 1405
constructivism, definition of 1307
constructivist learning theory 1335
constructivist learning tools 1309
constructivist pedagogy 2286
consumer initial trust 2235
consumer search behavior 1673
content 169
content adaptation 403
content and discourse analysis 299
content management 996
content management systems (CMS) 1378
content management, definition of 996
content personalization for mobile interfaces 992
content production 1368
content standards 1505
content validity 1961
content, definition of 226
content-based filtering 1630
content-centric learning unit 1361
context 20, 232, 236, 393, 630, 636, 639, 930, 945, 946, 947, 948, 996
context adaptation 2639
context information 992
context information acquisition and trust service (CATS) 2426
context information, definition of 261
context metadata 993–995
context metadata, definition of 996
context of performance 1925
context reasoning 237
context representation 237
context, acquisition of 237
context, definition of 930, 996, 2626
context, five W's of 930
context-aware 636
context-aware adjustment 639
context-aware animations 636
context-aware applications 236 236
context-aware computing 1019
context-aware language model 1005
context-aware services 21, 328
context-awareness 112
context-awareness 584, 1018, 1031
context-awareness, location-awareness 1018, 1020
contextual inquiry 939, 981
contraction 2451
contracts 239
convergence 2324
convergent validity 862, 1962
convergent validity, tests for 1962
conversational grounding 1805, 1816
cookies 1086, 1630
cooperation 2316
cooperative design 677
core authors, definition of 314
core concepts, definition of 314
core journals, definition of 314
corporate creativity, principles of 2112
cost 1511
course management system, definition of 1870
coword analyses, definition of 314
Cramer Rao bound (CRB) 1175, 1192, 1193, 1194, 1195
Creative Commons 1251, 1255, 1258, 1261
creative redundancy 2113
creative technology 2267
creativity 377
criteria reduction 632
criterion-related validity 1962
criterion-related validity, test for 1963
critical information systems research 1881
critical success factors 1945, 1949, 2133
crossbow platform 2724, 2732
cross-cultural information systems 1755
cross-cultural psychology 133
cross-cultural studies 139, 1818–1822
cross-culture differences 2644
cross-media adaptation 2640
cryptographic primitives 340
cryptographic tools 333, 340
cryptographic tools, overview of 340
cryptography 327
cryptography in UC, limitations of 343
cryptography in UC, potential in 343
cryptography, asymmetric 327
cryptography, asymmetric, definition of 331
cryptography, limitations of 343
cryptography, potential of 343
cryptosystems 340
cryptosystems, asymmetric 341
cultural barriers 1772
cultural diversity 2329
cultural expression 2332
cultural homogenization 2329
cultural knowledge 2259
cultural literacy 2330
cultural models 2655, 2657, 2659, 2660, 2663, 2664, 2669, 2671, 2674
cultural probes 938
cultural responsibility of IT professionals 2338
cultural shift 2325
cultural studies perspective 981
culture 440, 987, 1056, 1063, 1571, 1817, 1820
culture and learning 1350
culture theory 132, 136
culture, national 143
current goal states (CGS) 954
current learning goal states (CLGS) 955
curriculum resources 1505
customer delight 1655
customer delight, definition of 1655
customer interfaces 394, 398
customer relationship management (CRM) 1649
customer relationship management (CRM) 213
Index

customer satisfaction 1655
customer satisfaction, definition of 1655
customizable products 168
customization 1630
customization, mass 176
cyber-societies 2316
cyberspace 1823, 2337

D
data analysis 793, 2649
data analysis and visualization 1910
data collection 1573, 1909, 2095, 2167, 2363
data collection and analysis 1615
data mining 873, 874, 874–898, 876, 879, 880, 890, 893–896, 893–898, 1086
data mining algorithms 872, 893
data mining association rules 876–898
data mining systems 893, 893–898
data processing 2169
data smoothing 1599
data to information (D2I) environment 894
data, origin authentication 342
data, source federation 116
database 401
dataflow 636, 639, 1939
dataflow architectures 631
dataflow networks 636
DATMap 2269
deception models 1972
deceptive communication, effectiveness of 1970
decision-making 2316
decision-making communication protocols 1858
decision-making satisfaction 1960
decision-support systems and tools 176
declarative applications in immersive sensor networks (DAIS) 2708, 2722, 2723, 2724, 2726, 2728, 2729, 2730, 2731
decryption 340
dedicated hosting 2610
Defense Advanced Research Projects Agency (DARPA) 268
define task activity 953
dedeindividuation 1745, 1752
delegation 325
delicious 2593
delight 1652
della Francesca, Piero 492, 493, 494, 495, 502, 503, 505, 507, 513, 515
denial-of-service (DoS) attacks 337, 348
dependence 2666
description logics 1101
descriptive modelling 908
design efficiency 2108
design guidelines 486
design optimization 631
design principles 988
desktop multimedia 95
detection confidence 1987
determinism 1818–1819, 2656
developing countries 1591
developing country municipality 1608
development team 440
device federation 114
device identification engine 275
device mobility during a usability evaluation, definition of 2589
device-agnostic 2711
devices, ambient, orbs 271
devices, Blackberry 270, 275
devices, Bluetooth 231, 271, 286, 996
devices, Bluetooth, definition of 996
devices, Bluetooth, service discovery in 287
devices, cell phones 234, 269
devices, embedded 254
devices, embedded, definition of 261
devices, laptops 270
devices, mobile 254, 259, 262
devices, mobile, definition of 261
devices, mobile, input modalities for 111
devices, mobile, output modalities for 113
devices, pagers 269
devices, personal computers (PCs) 266
devices, personal digital assistants (PDAs) 168, 222, 263, 271, 275, 1087, 1639, 1649, 2106, 2286
devices, pervasive 269
devices, portable 259
devices, tablet PCs 270, 1442
dialog control 636
dialogue, interactive 215
didactic principles 988
didactics 987
differentialphase-shift keying (DPSK) 1038
Diffie-Hellman (DH) protocol 343
digital audio player, definition of 2540
digital autonomous cultural objects (DACOs) 499, 500, 501, 503, 509
digital communication networks 1059
digital communication technology 2604
digital content culture 1063
digital context culture 1065
digital cultures 1061, 1062
digital device 2593
Index

digital distribution culture 1064
digital divide 1511
digital e-culture 2324
digital habitats 552
digital imaging trek 2590, 2597 2591, 2598, 2600
digital inclusion 2542, 2543, 2546, 2547, 2551, 2552, 2553
digital information 2327
digital ink 1446
digital literacy 1490
digital media 2327
digital pen 1440
digital photography 2600
digital rights management (DRM) 1065
digital signature algorithm (DSA) 342
digital signatures 327, 342
digital subscriber lines (DSLs) 265
digital technologies 95
digital technology architecture 2602
digital television (DTV) 516
digital Web culture 2543
digitally enabled traveller 2590 2591, 2592, 2593, 2597, 2604
direct interaction 2640
direct sequence spread spectrum (DSSS) 1039, 1043
directed dialog 1222
directly related actions 2402
direct-manipulation 982
directory 1677
directory, definition of 1677
disability categories 2551
disability, categories of 2550
disambiguation, mutual 96
discourse analysis 2299
discourses of technology and gender 2300
discovery messages 235
discovery messages, DiscoveryHealth.com 826
discriminant validity 862
display-based problem solving 467
dispositional perspective 1817
distance communication, 2327
distance education 1864, 2278
distance education technologies 2278
distance learning 1865
distributed cognition 912, 935
distributed cognition model 932, 933, 935, 936, 942, 944, 946, 948
distributed communities of practice (DCoPs) 979, 980, 982
distributed component 640
distributed deception 1970

distributed Digital Incunabula Library 498
distributed frameworks 631
distributed knowledge 1102
distributed provision 285
distributed service provision, example of 285
distributed systems 14
distributed wearable augmented reality framework (DWARF) 639, 642
distributed work 2028
DIVE 636
diversity feature 1777
doing gender 2036
domain knowledge representation 742
domain ontology acquisition 735
domains of instructional objectives, definition of 2555
domain-specific ontologies 244
dynamic host configuration protocol (DHCP) 234, 286
dynamic interface 220
dynamic queries, processing 2726
dynamically discoverable 281

E
e-action 2395
early adopters 1204
ease of use (EOU) 1651, 2334
e-based task artefacts 1717
e-books 168
e-business, human face of 2235
e-collaboration and satisfaction 1764
e-collaboration and trust 1763
e-collaboration, academic Weblogs 1291
ecological context, definition of 1250
ecological interface design (EID) 486
ecological task analysis 909
e-commerce 1649
e-commerce 177
e-commerce customer relationship management (e-CRM) 1995
e-commerce Web sites 2235
e-community 1632
economically developed countries (EDCs) 1856
economy 2297, 2298, 2300, 2308
e-culture 2325, 2331
e-culture literacy 2325
education and learning, innovative technologies for 1308
education and training 1945
education portals, Weblogs 1298
educational research 310
Index

edutainment 1205, 1380
efficacy of an evaluation method, definition of 2589
elder care 2026
e-learning 1350, 1354, 1355–1357, 1361, 1366, 1368, 1375–1385, 1379, 1435
e-learning content development applications 1368
e-learning environments 168
e-learning initiatives 1777
e-learning with social software 1402
e-learning, tools and approaches to 1377
electronic Beowulf Project 492, 493, 515
electronic computer-mediated communication 1563
electronic guidebook research project 1641
electronic product code (EPC) 353
electronic store 2396
electronic toolkit 1435
electronic whiteboard 1441, 1450
ELENA project 1634
e-lifelong-learning experience 961
eccentric curves (EC) 341
e-mail 1819
e-mail security 327
e-mailing intelligence 2164
embodied conversational agents 2640
embodied interaction 1017, 1020, 1031
embodied metaphor 1822–1826
empathy 440
encoding, definition of 226
encryption, hybrid 343
enculturation 1827
end user 440
end user license agreement (EULA) 2682
end users, definition of 32, 194
enduring involvement (EI) 2001
end-user applications 2711
end-user experience 1771
energy restrictions 233
enterprise reconfiguration dynamics 1055
enterprise resource planning (ERP) 2132
enterprise resource planning (ERP) implementation, grounded model for 2135
enterprise resource planning (ERP) system implementation 2132
entity 996
entity relationship diagrams 1939
entity, definition of 996
environment 793
environment data 2169
e-passports 353
e-passports, security measures of 354
equal opportunities commission (EOC), definition of 1589
equalization phenomenon 1752
equipment damage 1512
equitable use 2110
ergonomics 488
error and cognition 2548
error handling techniques 1009
error models 2549
e-speak 1088
essentialism 1760, 2315
essentialist theory 1754
technology 2325
ethics 436
ethnographic research 2658, 2673
ethnography 439, 677, 937, 1786, 2670
ethnomethodology 439, 2656, 2660, 2671, 2672
ethno-narrative perspective 981
EuroFlora guide 1241
European Monitoring Center for Drugs and Drug Addiction (EMCDDA) 1681
evaluation criteria 310
event scenes 1872
event service 241
event-based systems 240
eventing 286
events 1871, 1872, 1873
everyday life, phenomenological approach to 2655
everyday, view of 1875
evolutionary programming (EP) 379
evolutionary strategies (ESs) 379
executive information systems (EIS) 147
existing business modeling language 2390
expectance theory 1655
expectance theory, definition of 1655
expectancy 1650
expectancy theory 1650
expectation-states theory 1752
Expedia.com 826
experience design 433, 442
experience-based learning 2217
experiential learning, IT implementation as 2218
explicit specification 2150
exploration depth, analysis of 2250
exploratorium 1641
external cognition 467
external memory 467
external stakeholder integration 2101
external technical integration 2100
externalization of information 466
eye tracking 630, 636, 639

F
Facebook 2593
face-to-face (FtF) 1745, 1850, 1851
face-to-face (FtF) communication 2120, 2123
face-to-face (FtF) communication elements 1550
facial expressions 96
facility management 273
family education, socio-cultural changes in 2664
family interfering with work (FIW) 2025
family pattern 136
family-work contrast barriers 2033
Far East cultures 126
FataMorgana 630, 633, 641
federative library (FedLib) 2192, 2193, 2204, 2205, 2206
FedLib, smart office environment for 2206
feedback management 1022
female only (FO) 1553
feminine cultures 147
femininity 1571, 2030
feminism 2299, 2305
feminist gaze 2299
feminist methodology 2268
feminist research methodologies 2268
feminist theory 2297, 2298, 2306
final exam grade, definition of 1870
Finnish innovation system 2041
FireFly 263
first mile-last mile 265
fission 1214
flexibility 96, 224, 2334
Flickr 2594, 2602
flow charts 1939
flow theory 1998
FoodTV.com 826
formal e-learning 1634
formality domain indicator 1824
formation phase 1920
formation phase, challenges in 1924
formation, content of 1921
formation, context of 1920
formation, process of 1923
formative modelling 908
formulate procedure activity 953
forums 1380–1381
frames 97
free and open source software (FOSS) 1459, 1467
frequency-hopping spread spectrum (FHSS) 1038, 1043
full disclosure 1842
full function device (FFD) 1039
functional complexity 1998
fundamental scalability challenges 2147
fusion 1214

G
Gandhi, Mahatma 1780, 1784
gatekeeping 2540
gatekeeping, definition of 2540
Gaussian frequency shift keying (GFSK) 1038
GCE A level, definition of 1589
GCSE, definition of 1589
gender 2259, 2260, 2261, 2262, 2263, 2296, 2297, 2298, 2301, 2302, 2303, 2304, 2305
gender and computing 1583
gender and information technology 1591
gender and technology 2297, 2306
gender computing gap 1737, 1744
gender differences 26, 2287
gender differences in spatial abilities 2475
gender roles 2277
gender segregation 2296, 2297, 2306
gender stereotypes 2036
gender, human-computer interaction 2474
gender-appropriate 1587
gender-based CMC style 1565
gender-based differences 2481
gendered body 2264
gendered place experience 2264
gender-inappropriate 1587
genereal event notification architecture (GENA) 242
general meta-choice console (GMCC) 2684
general packet radio system (GPRS) 234
genetic algorithms (GAs) 379
geographical adaptive fidelity (GAF) algorithm 1174, 1177, 1178
geographical proximity 2045
GeoGRID algorithm 1174, 1177
GIGA 1224
glass ceiling 2036
global communication ethics 2454
global community 1565
global information ethics 2450, 2454
global information ethics, advantages 2455
global information ethics, cost of 2457
global information technology (GIT) 1849, 1850, 1858
Index

global information technology (GIT) infrastructure 1853
global knowledge economy 2296, 2297, 2300
global positioning system (GPS) 1174, 1178, 1179, 1192, 1193, 1196, 1198
global positioning system (GPS), definition of 2571
global travel industry 2593
global village 600
globalization 132, 1066, 1435, 2047, 2328
GLUE 1104
GOAFR+ algorithm 1174, 1176
goal states (GS) 954
goals, operators, methods, and selection (GOMS) model 2548, 2554, 2555
goals, operators, methods, and selection (GOMS) model, definition of 2555
good degree, definition of 1589
Google Talk 2612
governance knowledge network (GKN) 980
governance knowledge network (GKN) OLC 980
governance knowledge network (GKN) prototype 980
graph reading, modelling 1163
graphical user interfaces (GUIs) 1389, 2565
graphical user interfaces (GUIs) 434, 455, 659, 660, 1017, 1018, 1022, 1023, 1030
graphical user interfaces (GUIs) testing 659
graphical user interfaces (UI) class diagram 412–428
graphical user interfaces (UI) prototypes 416–428
graphics tablet 1440
GRID algorithm 1174, 1176, 1177, 1197, 1198
grid computing 254, 255, 258
grid perspective 258
grid services, definition of 261
grids, as test bed of pervasive applications 257
grids, mobile/wireless 258
grids, to collect data from sensor networks 257
grids, usage of 256
grounded theory 2134
grounding business interaction 2082
grounding business interaction models 2071
GroupWare task analysis (GTA) 908
Gujarat 1781

H
hackers 324, 1737
handwriting recognition (HR) 1216
haptic output 96
hash algorithm 341
hash functions 341
head up display (HUD) 637, 638, 639
head-mounted display 629, 633
health information 1687
health, managing 1687
healthcare IT project 1592
helpfulness 2675
heterogeneous group 1752
heuristic evaluation 455, 706–707
heuristic evaluation approach 455
hidden Markov model (HMM), definition of 2571
hierarchical structure 136
hierarchical task analysis 901, 907
hierarchy 139
hierarchy/egalitarianism (H/E) 139
high context/low context of communication 132, 142
higher education institutions (HEI) 2132
higher education level, computer science at 1584
Higher Educational Statistics Agency (HESA), definition of 1590
higher group communication style (HCS) 1563
higher group communication style (HCS) communication 1556
Hindi 1780
HomeRF 263
homogeneous group 1752
horizontal gender segregation 2036
host discovery 234
hosting 2610
human actions 2392
human computer interaction (HCI) 1385, 1388
human computer interaction (HCI) 1435, 1436–1437, 1656, 1656–1659
human computer interaction (HCI) in ETD, role of 1680
human computer interaction (HCI) mediating tools 1393
human computer interaction (HCI) models 453, 459
human computer interaction (HCI) theory 450
human computer interaction (HCI), definition of 1250
human culture 1941
human element 1632
human factor design principles 2107
human factors 60, 68, 631
human factors engineering 2111
human factors engineering, definition of 2116
human factors environments 2111
human factors for networked and virtual organizations 2106
<table>
<thead>
<tr>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>human factors taxonomy 2109</td>
</tr>
<tr>
<td>human factors taxonomy, developing a 2109</td>
</tr>
<tr>
<td>human factors, challenges of 1939</td>
</tr>
<tr>
<td>human information interaction, evolution of 2693</td>
</tr>
<tr>
<td>human inner world 2392</td>
</tr>
<tr>
<td>human interaction 197</td>
</tr>
<tr>
<td>human needs 436</td>
</tr>
<tr>
<td>human perception processes, mechanisms of 39</td>
</tr>
<tr>
<td>human-centered design 2106</td>
</tr>
<tr>
<td>human-centered design 2550, 2629</td>
</tr>
<tr>
<td>human-centered design process 2108</td>
</tr>
<tr>
<td>human-centered information spaces 2108</td>
</tr>
<tr>
<td>human-centered interface design 2549, 2555</td>
</tr>
<tr>
<td>human-centered interface design, definition of 2555</td>
</tr>
<tr>
<td>human-centered technologies 677</td>
</tr>
<tr>
<td>human-centric approach 377</td>
</tr>
<tr>
<td>human-computer communication (HCC) 704</td>
</tr>
<tr>
<td>human-computer interaction (HCI) 62, 68, 73, 80–94, 194, 212, 262, 324, 1769, 1998, 2474, 2548, 2675</td>
</tr>
<tr>
<td>human-computer interaction (HCl) and gender 25, 32</td>
</tr>
<tr>
<td>human-computer interaction (HCl) defining evaluation targets 81</td>
</tr>
<tr>
<td>human-computer interaction (HCl) designing an evaluation protocol 88–89</td>
</tr>
<tr>
<td>human-computer interaction (HCl) making sense of human activity 85–86</td>
</tr>
<tr>
<td>human-computer interaction (HCl) referent models 83</td>
</tr>
<tr>
<td>human-computer interaction (HCl), benefits of 224</td>
</tr>
<tr>
<td>human-computer interaction (HCl), modes of 262</td>
</tr>
<tr>
<td>human-computer interaction (HCl), security for 323</td>
</tr>
<tr>
<td>human-computer interaction (HCl), task ontology-based 950</td>
</tr>
<tr>
<td>human-factors engineering (HFE) 488</td>
</tr>
<tr>
<td>human-friendly interface 951</td>
</tr>
<tr>
<td>HV/collectivism 1892</td>
</tr>
<tr>
<td>HV/individualism 1892</td>
</tr>
<tr>
<td>hybrid authoring 631</td>
</tr>
<tr>
<td>hybridization 2452</td>
</tr>
<tr>
<td>Hypercard 1467</td>
</tr>
<tr>
<td>hyper-parenting 2655, 2658, 2660, 2664</td>
</tr>
<tr>
<td>hyperspatial 2332</td>
</tr>
<tr>
<td>hypertext markup language (HTML) 270</td>
</tr>
<tr>
<td>hypertext transfer protocol (HTTP) 269</td>
</tr>
<tr>
<td>physical context 236</td>
</tr>
</tbody>
</table>

I

i-actions 2393
IBM 275

ICOM/CIDOC standards 510, 514
icon search, modelling 1166
ID-based cryptography 2428
idea evaluation stage 1855
idea generation stage 1855
IDEF0 489, 491, 492, 503, 504, 505, 506, 511, 512
identify activity 953
identity 2259, 2260, 2261, 2263
immersive authoring 631
immersive environments 2730
immersive networks 2715
immersive sensor networks 2709, 2710, 2711, 2712, 2713, 2714, 2715, 2730
immersive visual programming environment 636
i-Mode 1649
implementation neutrality 281
implicit-theoretical research 2309
impolite computing 2680
impolite effect 2684
impression formation 1837, 1838, 1839, 1840, 1841, 1842, 1843, 1844, 1845, 1846, 1847, 1848
impression formation, definition of 1848
impression management 1841, 1848
impression management, definition of 1848
In touch project 1639
inappropriate use of technology 1512
in-car user-interfaces 532–551
in-car user-interfaces case study: vehicle navigation systems 544
in-car user-interfaces design and evaluation 537
in-car user-interfaces design and evaluation 15 second rule 543
in-car user-interfaces design and evaluation field trials 539
in-car user-interfaces design and evaluation keystroke level model (KLM) 543
in-car user-interfaces design and evaluation lane change task 543
in-car user-interfaces design and evaluation peripheral detection task 542
in-car user-interfaces design and evaluation road trials 540
in-car user-interfaces design and evaluation simulator trials 540
in-car user-interfaces human-centered design process tasks 535
in-car user-interfaces human-centered design process, equipment 536
in-car user-interfaces, human-centered design process 534
in-car user-interfaces, human-centered design process, environments 537
Index

in-car user-interfaces, human-centered design process, users 535
in-car user-interfaces, types of in-car computing systems 533
inclusion 2542, 2543, 2546, 2547, 2551, 2552, 2553
inclusive pedagogy 2286
inconsistent tagging 1600
indefinite actions 2394
India, governmental agencies 1779–1781
indirectly related actions 2402
individual difference theory 1754, 1756
individual differences 2311
individual differences theory of gender and IT 1760, 2315
individual interest 1633
individual motivation and competence 2183
individualism 1571
individualism/collectivism (I/C) 132, 137, 139, 1816
individuals with disabilities 2543, 2547, 2552, 2553, 2555
informal e-learning 1634
informatics (I) 1744
information access 1506
information age 2112
information and communication cultures 1058
information and communication technologies (ICTs) 1307, 1388, 1436, 1594, 1769, 1849, 2296, 2297, 2298, 2307, 2590, 2602, 2656, 2658
information and communication technologies (ICTs) and the Global Knowledge Economy 2297
information and communication technologies (ICTs) services 2602
information and communication technologies (ICTs), definition of 1307, 1594
information and communication technologies (ICTs), in classrooms 1437
information and communication technologies (ICTs), infrastructure 2602, 2603
information appliance 2333
information assistants 2675
information behavior 2336
information communications systems 2325
information customization (IC) systems 1120
information design, definition of 2116
information dissemination 2022
information economy 2331
information ethics 154, 155, 156, 157, 161, 163, 164, 165, 166, 167
information extraction (IE) 1080
information foraging on the Web 1159
information gathering 1675
information infrastructure 2193, 2198, 2199, 2201
information presentation 1959
information quality 1958
information retrieval (IR) 1080, 1105
information richness 2122
information richness, definition of 2189
information seeking 825
information society 2324, 2329
information systems (IS) 133, 605, 617, 619, 620, 791, 2052, 2390
information systems (IS) designers 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 616, 617
information systems (IS) development 605, 619, 620, 645
information systems (IS) research methodology 1871
information technology (IT) 132, 245, 333, 1753, 2298, 2300, 2301, 2304, 2306, 2309, 2327
information technology (IT) artefacts 2341
information technology (IT) management 2338
information technology (IT) outsourcing 1754
information technology (IT) security, objectives of 337
information technology (IT), adoption of 143
information technology (IT), future of 2469
Information Technology Workforce (ITWF) project 1738
information visualization, definition of 1685
information-access 262
informational environmentalism 2457
informational habitat 556
informational ontology 2456
informational politeness 2679
information-communications systems 2324
information-communications systems convergence paradigm 2324
information-friendly environment 1060
information-friendly societies 1058
information-generation 262
information-matter-energy model 35
infrared light 352
ingenious project 1636
inhabited knowledge space (IKS) 701
inherent differences 2310
in-network aggregation 2712
in-network processing 2711
in-network query resolution 2731
input and output federation 115
input-process-output models 1816
instant communication 272
instant messaging (IM) 1310, 1561, 2611
instructional objectives, domains of 2546, 2555
instructional technology (IT)  1744
instructor-led training (ILT)  1354
instrumental personalization  853
instrumented environment  2709
insufficient-theoretical research  2309
integration  2092
integration manager  1214
integration schemes  2431
integration support in RUP  2093
integration types  2093
integration, managerial challenges of  2100
intellectual property rights (IPR)  1065
intellectual turning points, definition of  1686
intelligence  2641
intelligence, concept of  553
intelligent agent paradigm  2621
intelligent agent, definition of  2626
intelligent interaction  2640
intelligent learning  900
intelligent networks  2327
intelligent user interfaces (IUIs)  2614, 2627, 2638
intelligent user interfaces (IUIs), definition of  2626
intelligent virtual sensors  2718
intentional collectives  2398
intentionality  2656, 2673
interaction  917
interaction design  64
interaction mode  275
interaction situations  2685
interaction techniques  631
interaction, lean mode of  1752
interactions, ad hoc  335
interactive behaviour modelling  1154–1172
interactive evolutionary computation (IEC)  377
interactive evolutionary design systems (IEDSs)  377, 378
interactive LCD panel  1440
interactive multimedia exhibition  2269
interactive systems design  681
interactive tools  636
interactive voice response (IVR)  901
interactivity  1506
interactivity design, definition of  2116
inter-class constraints  632
intercultural communication  1787, 1795
interface  394
interface agents  1070
interface design  397
interface satisfaction  1960
interfaces  69
interfaces, design of  67, 205
interfaces, interactive  196
interfaces, requirements for  202
interferometric ranging  1175, 1178, 1180, 1181, 1182, 1189, 1192, 1194, 1195, 1196
intermittent resources, definition of  261
internal ERP knowledge  2142
internal stakeholder integration  2100
internal technical integration  2100
internalization of information  466
International Civil Aviation Organization (ICAO)  354
International Society for Technology in Education (ISTE)  1505
Internet access, mobile  169
Internet activities  1823
Internet protocol (IP)  268
Internet protocol (IP) phones  270, 271
Internet radio, definition of  2540
Internet relay chat (IRC)  1549
Internet service providers (ISPs)  266
Internet shopping site personalization preferences  2355
Internet usage  2365
Internet-friendly classrooms  1202
interoperability  1091, 1098
interpersonal cues  2241
interpret activity  953
interruptability  22
intrinsic motivation  1819
invisible culture  2330
invisible e-culture  2333
invisible e-technologies’ interaction  2333
invisible networks  2467
iPod  1223
iPod  2613
ISD process for social learning  2061
ISO 9241-11 standard  361, 362, 363, 364, 367
Istanbul-Maslak  2047
IT/computer professional, definition of  1590
Italian Public Administration  791
ITC cross-cultural literature  2644
iterative information system development projects  2090
IT-supported learning and teaching, consequences for  1407
iTunes Music Store  1223
J
jam sessions  636
Java 2 Micro Edition (J2ME)  270
Java Server Pages (JSP)  811
Index

Java Virtual Machine 270
JavaBean 812
Jini 243, 286, 1088
journals 299
journals, education-related 299

K
K-16 2286
Kaiser-Meyer-Olkin (KMO) 520
karaoke 1224
Kelly, Kevin 4, 5, 7
Kerala Government 1781
key exchange 343
key scheduling 343
kinetic user interfaces (KUI) 1015, 1021
kinetic user interfaces (KUI) components 1029
kinetic user interfaces (KUI) conceptual taxonomy 1025
kinetic user interfaces (KUI) interaction patterns 1023
kinetic user interfaces (KUI) middleware 1028
kinetic user interfaces (KUI) middleware, enabling technology 1030
kinetic user interfaces (KUI) model 1025
kinetic user interfaces (KUI) toolkit, sketch of 1029
kinetic user interfaces (KUI) vs. GUI 1022
kinetic user interfaces (KUI) widget (Kuidgets) 1025, 1026, 1027, 1029, 1030
kinetic user interfaces (KUI), interaction patterns 1017, 1023, 1024, 1029, 1030
kinetic user interfaces (KUI), interaction patterns, continuous tracking 1023, 1031
kinetic user interfaces (KUI)-enabled scenarios 1024
kinetic-aware systems 1019
kinetic-awareness 1018
knowledge age 2112
knowledge age, principles of corporate creativity 2112
knowledge city 2047
knowledge creation 980, 982
knowledge discovery (KD) 872, 873, 873–898
knowledge discovery in databases (KDD) 1081, 1086
knowledge economy 2296, 2297, 2300, 2307
knowledge engineering 678, 2274
knowledge industry 2047
knowledge management systems (KMS) 1379
knowledge management, definition of 2189
knowledge networking 2014
knowledge of baseline behavior 1986
knowledge sharing, British Council 1905
knowledge sharing, promoting 1908
knowledge society 2051
knowledge systems, formal 40
knowledge work 2028
knowledge, nature of 1906
knowledge-based systems (KBS) 704
Korean Information Infrastructure (KII) 1202

L
Label Mobil 1225
lack of participation 2304
language 439, 1817
language action perspective (LAP) 2390
language action perspective (LAP) community 582
language labs 1364, 1364–1366
language learning 1357, 1357–1366
language learning, traditional and innovative 1359
language understanding 1006
language/action (L/A) community 2076
layered reference model of the brain (LRMB) 685, 686, 688, 696, 697
leadership 1948
learner profile 1046
learner profile data, acquiring the 1048
learner satisfaction 1356
learning 1864
learning and IT implementation 2216
learning as acting 2224
learning as conceptualizing 2225
learning as planning 2227
learning as reflecting 2224
learning autonomy 1356
learning communities 1257–1258
learning communities, definition of 1307
learning community 1307
learning content management system (LCMS) 1379
learning ecologies 1392
learning efficiency 1510
learning environments 1389
learning goal states (LGS) 954
learning objects (LOs) 2286
learning objects (LOs) 246
learning platform 319
learning platform, definition of 322
learning portals 1384
learning processes, role of 2228
learning unit 1362
legacy help desk 1707
legitimate interactions, politeness support of 2677
less economically developed countries (LEDCs) 1855
Index

lexical conceptual graph (LCG) 1103
lexical mapping theory 2501
lexicalized HMMs 1595, 1597
library and information science/studies (LIS) 1744
lifelong learners 1639
lifelong learning 1379
Likert scale, definition of 2589
linguistic communication 1848
linguistic communication, definition of 1848
linguistics 1817
link manager protocol (LMP) 1038
liquid crystal display (LCD) 525
LiveJournal 1488
local area network (LAN) 262
local area network (LAN) technologies 265
localization 1173, 1174, 1175, 1176, 1177, 1178, 1179, 1180, 1181, 1182, 1183, 1184, 1185, 1186, 1187, 1188, 1189, 1190, 1191, 1192, 1193, 1194, 1195, 1196, 1197, 1198
location 21, 996
location aided routing (LAR) algorithm 1174, 1176, 1177, 1196, 1198
location aware, definition of 2571
location paradigm 21
location sensing 21
location, definition of 996
location-aware computing 1237
location-aware computing, definition of 1250
location-aware services 1019
location-based multicast (LBM) algorithm 1174, 1177
location-based services 273
locative inversion in Chinese 2512
locus of control 1819
logical link control and adaptation protocol (L2CAP) 1038
London Science Museum 1639
long-term relationship 1826
loose coupling 281
low- vs. high-context of communication 1816
lurker 1823

M

Mac OS 264
machine learning 378, 1120
MADE application 1238
MADE architecture 1238
male dominance 1748
male only (MO) 1553
management information systems (MIS) 145, 364, 365, 373
management information systems (MIS), computer-based 145
management support 2137
managerial positions 2031
man-in-the-middle (MITM) attacks 338
manipulation check 863
mapping 1103
maps 178
marginal communities 2015
marketing 1677
marketing, definition of 1677
masculinity 1571
masculinity/femininity (MAS) 132, 137, 140
Maslow’s hierarchy of needs 1351
mathematical proof, social nature of 2437
maturation effects 2108
meaning, social negotiation of 1820
measurement model 1894
media affordances 1816
media choice, definition of 2189
media design, definition of 2116
media gap 1356
media richness 1655
media richness theory (MRT) 1550, 2122
media richness theory (MRT), classification criteria of 2122
media richness, definition of 1655
media services language (MSL) 1235
mediating tools 1388
medical professionals 1594
medical professionals, definition of 1594
medium access control (MAC) 1037, 1043
mental context 993
mental models 196
mesh radio 267
message authentication codes (MACs) 342
message authentication codes (MACs), hash-based (HMACs) 342
message board 1823
message format 1651
metacognition, definition of 960
metacommunication 450
metadata 2150
meta-model 2392
meta-model of SIP 2397
meta-model socio-instrumental pragmatism 2391
meta-object-facility (MOF) 762, 765, 769, 789
metaqueries 875–876, 876
metaqueries semantics 885, 885–886
methodology 659, 660, 662, 664, 665, 666, 670, 675, 676
methods of computer-mediated communication, definition of 1870
metrics 2701
microcontroller 268
micro-electrical-mechanical systems (MEMS) 1136
microelectronics 279
microphone array processing 1009
micro-planning 903
Microsoft 275
Microsoft FrontPage 2605
Microsoft Network (MSN) 2612
Microsoft Publishing 2605
middleware 1015, 1021, 1023, 1025, 1028, 1029, 1030, 1031, 2708, 2709, 2711, 2712, 2713, 2714, 2715, 2718, 2720, 2722, 2724, 2727, 2728, 2729, 2730, 2731, 2732, 2734
MIDI 1225
military standards for software 2441
MINERVA 508
mining task 889
mixed gender (MIX) 1553
mixed initiative multimodal dialog 1010
mobile access to information and applications 273
mobile ad hoc network (MANET) 1174, 1176, 1177
mobile commerce (m-commerce) 394, 1655
mobile commerce (m-commerce), customer satisfaction 1649
mobile commerce (m-commerce), definition of 1655
mobile commerce (m-commerce), personalization 1649
mobile communication technologies 2655, 2658
mobile computing 1015, 1022, 1034, 2614
mobile computing, definition of 2626
mobile device, definition of 2571
mobile devices 992
mobile devices, voice-enabled user interfaces for 2556
mobile digital technology 2591
mobile devices, content personalization for 992
mobile multimedia (M3) 394
mobile multimedia (M3) communication 400
mobile multimedia (M3) presentation 400
mobile phones 435, 1213, 2655, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2666, 2667, 2669, 2670, 2671, 2673
mobile phones, dialogical use of 2669
mobile phones, diffusion and appropriation of 2658
mobile phones, negotiation tool 2665
mobile phones, usage of 1569
mobile salesman 334
mobile technology 997
mobile tourist guide, definition of 1250
mobile ubiquitous computing systems 1015
mobile user interfaces 2573
MOBIlearn 677
mobility 233, 394
mobility, degree in 234
modalities 95
modalities, combinations of 96
modalities, replacement 106
modality fusion 1006
modality fusion, definition of 1014
modality switching 1004
modality, definition of 1014
model human processor model 932
model of the brain, cognitive informatics (CI) 37
model of the brain, layered reference 37
model view controller 714
model-based user interface management systems (MB-UIMSSs) 761
modeling/designing 1940
modern classroom 1436
modification detection code (MDC) 341
modifier/extender 2029
modulation 1043
monitor knowledge state 955
monitor learning plan 955
monitoring tool 630, 635
Moore’s Law 2592
moral economy 2657, 2660, 2669, 2672
moral interactions, sources and targets 2457
morpholexical processes 2515
morphosyntactic processes 2515
MOSQUITO architecture 2426
MOSQUITO components 2426
MOSQUITO framework 2427, 2429
MOSQUITO project 2430, 2424, 2426
MOSQUITO scenario 2427
motes 268
motion as an input modality 1018, 1022
motivation 27, 29, 1946
motivation during adoption 1947
motivation factor 1947
motivation/attitude-driven behavior (MADB) model 690, 691
mouse clicks 815
MP3, definition of 2540
multi-agent system for academic advising (MA-SACAD) 1118–1133
multi-agent system for academic advising (MA-SACAD), benefits and limits 1130
multi-agent system for academic advising (MA-SACAD), system architecture 1124
multi-agent system, definition of 2626
multi-agent-based solution 1124
multilateralation 1179, 1183, 1185, 1186, 1197, 1198
multilayered evaluation approach 2576
multi-layered evaluation approach, definition of 2589
multiliteracies 1388
multimedia 393, 1258
multimedia applications 169
multimedia communication 401
multimedia content 393
multimedia database 401
multimedia language lab configuration 1365
multimedia message 1652
multimedia messaging (MMS) 992
multimedia personalization 394
multimodal and federated interaction 102
multimodal fusion 104
multimodal interaction (MMI) 95, 103, 274, 1222, 2640
multimodal interaction in context-adaptive systems (MICA) 112
multimodal interface, definition of 2571
multimodal output 110
multimodal user interface (MUI) 997, 998, 1213–1222
multimodal user interface (MUI), definition of 1014
multimodality 102, 1388
multimodality and accessibility 109
multimodality and security 108
multimodality for mobile devices 111
multimodality on the desktop 105
multimodality, basics of 102
multimodality, different forms 103
multi-pedagogies 1391
multiple intelligences 1390
multiple objectives 386
multiple-view mode 528
multisensorial approach 1380
multisensoriality 1388, 1392
MundoCore 15, 16, 17
MUSIC.CO.JP 1224
mutation 380
My Yahoo! 171
MyExploratorium 1642
MySpace 1778

N

national multimedia resource centre (NMRC) 1780, 1785
National Museum of Ethnology 1643
natural environment 2392

natural intelligence (NI) 38
natural language 916
natural resistance to change 1942
navigation 169
navigation map 639
necessary culture change 1941
needs assessment, definition of 298
neo-symbiosis 2693, 2696
neo-symbiosis research agenda 2697
nesC 2713, 2724, 2728, 2733
netiquette 320
netiquette, definition of 322
NetStumbler 264
network paradigm 1853
networked communication 1849, 1851, 1855
networked communication work 1854
networking 2332
networking, ad hoc 20
networking, pervasive 233
networks, corporate 273
networks, mess 267
networks, metropolitan area (MANs) 265
networks, mobile communication 263
networks, neighborhood area (NANs) 264
networks, opportunistic, passive collaboration in 335
networks, peer-to-peer 268
networks, personal area (PANs) 263
networks, radiant 268
neural informatics (NeI) 38
new economy 2048, 2297, 2298, 2300, 2308
new media 2325, 2327
Ning 1778
non line of sight (NLOS) 265
non-reactive approach 2167
nonverbal communication 1848
nonverbal communication, definition of 1848
non-verbal cues 1820
Norman, Donald 4, 5, 7, 8
normative modelling 908
NTT DoCoMo 1224
number of generalized messages 1650, 1652

O

OA layer 277
object constraint language (OCL) 762, 789
object model 1939
object orientation, definition of 226
object-attribute-relation (OAR) model 685, 690, 696, 697
object-matching experiment 2477, 2483
object-oriented 1938
Index

object-oriented design 678
object-oriented software development processes 1937
object-positioning experiment 2478, 2485
object-resizing experiment 2478
object-role management (ORM) 583
objects 2392
observations 440, 457
observe activity 953
observer 241
ocation-based adaptation 21
offset-quadrature phase-shift keying (OQPSK) 1039
oiceXML 275
one-size-fits-all approach 224
online 1677
online collaborative writing 1262
online communities 1632
online effect 1707
online etiquette 2677
online exchange outcome 1819
online health information 1690
online health information, baby boomers and seniors 1689
online health resources 1691
online health seekers, sociodemographic characteristics of 1691
online journal 1484
online learning 2685
online learning communities (OLCs) 979, 980, 981, 982, 984, 988
online learning communities (OLCs), features of 983
online learning communities (OLCs), formal and informal 982
online politeness 2689
online shopping 1100
online social networks 1521, 1522, 1524
online task management system 1707
online trust and media cue 2241
online virtual environment 1745
online, definition of 1677
ontological categories 2391
ontological engineering 1099
ontologies 2144, 2145, 2146, 2148, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2160, 2161, 2162
ontologies, benefits of 2153
ontologies, limitations of 2155
ontology 1099
ontology engineering 735
ontology languages 2144, 2148, 2150, 2154, 2155, 2156, 2160
ontology, definition of 960
ontology-aware system 958
ontology-based personalization 748
OO processes, perception of 1950
open and distance learning (ODL) 1354, 1379
open information, building up 2203
open innovation 2193, 2197
open innovation, economic perspective 2197
open object information infrastructure (OOII) 2192, 2193, 2201, 2202, 2203, 2204, 2207
open source learning 1467
open source software (OSS) 645, 650, 656, 657, 658, 1459, 2337
open source, definition of 298
openness 2675
optimality-theoretic lexical mapping theory 2499
optimality-theoretic LMT 2507
organizational actions 2394, 2396
organizational change and human factors 1938
organizational change management 2137
organizational culture 2036, 2184
organizational learning 2216
organizational policies and support 2184
organizational taxonomies 2109
orientation, long-term 141
orthography 1820
OSGi 243
OT-LFG overview 2507
OT-LMT, comprehensive 2508
OT-LMT, illustration of 2510
OT-LMT, potential advantages of 2520
outcomes 1864
out-of-band channels 351
overconfidence 32
overhead projector 1438
overhead projector 1438
Oxford English Dictionary 1941

P

PAARTI 630, 641
paleolithic stone age effect 2473
paper sketches 459
paradigm model 2135
paradigm shift, definition of 1686
paralinguistic cues 1840, 1843, 1844, 1848
paralinguistic cues, definition of 1848
parallel coordinate box plot (PCBP) 386
parenting, cultural models of 2659
parenting, new cultural models of 2655, 2658, 2660, 2662, 2663, 2664, 2665, 2666, 2668, 2669, 2674
Pareto-frontier 388
Index

Paris-ile-de-France.com 826
participatory design 979, 982, 984
participatory prototyping 443
passwords 326
paternalism 136
patient monitoring 335
patient-controlled analgesia (PCA) 488
patterns 2431
pedagogical shortcomings 2283
pedagogical transformation 1656
peer culture 2658, 2668
perceived ease of use (PEOU) 142, 1651, 2294
perceived ease of use of general advertisements 1650
perceived usefulness (PU) 142, 1651, 1652, 2294
perceived usefulness (PU) of personalized message 1650
performance phase, challenges in 1928
performance, content of 1927
performance, process of 1927
perimeter security 2424
permission marketing 186
Perseus development corporation 1488
personal agenda 1642
personal identification number (PIN) 326
personal interests 821
personal knowledge publishing, definition of 1297
personal learning environments (PLEs) 967, 1407
personalization 168, 176, 179, 213, 394, 403, 807, 815, 1080, 1630, 1649, 1651, 1655
personalization and Web mining 2174
personalization design paradigms 855
personalization process 2166
personalization strategy 864
personalization strategy measures 861
personalization strategy, archetypes of 852
personalization systems 213
personalization systems, deployment of 212, 214
personalization systems, persistent 169
personalization techniques 1630
personalization, definition of 815, 1655
personalization, degree of 169
personalization, deployment of 216
personalization, effects of 855
personalization, interactive dialogue 215
personalization, persistent 169
personalization, techniques for 170, 213
personalized behavior 180
personalized communities 215
personalized content 2173
personalized metasearch systems 1083
personalized navigation 169
personalized output 2173
personalized page importance 1083
personalized presentation 2174
personalized services for collaboration tools 1049
personalized structure 2173
personalized virtual Web spaces 1635
personalized Web sites, user-centered evaluation of 177
persuasive agenda 2336
pervasive applications, test bed of 257
pervasive communication 234
pervasive computing (PC) 22, 228, 254, 255, 256, 262, 275, 276, 279–280, 281, 289, 1043, 2192, 2193, 2204, 2212, 2708, 2709, 2711, 2718, 2733
pervasive computing (PC) environments, service provision for 279
pervasive computing (PC) systems, development of 238
pervasive computing (PC) world 262
pervasive computing (PC), component-based approach in 239
pervasive computing (PC), context in 235
pervasive computing (PC), current and future trends in 244
pervasive computing (PC), definition of 289
pervasive computing (PC), event-based approach in 241
pervasive computing (PC), knowledge and learning management 245
pervasive computing (PC), middleware for 242
pervasive computing (PC), middleware solutions 243
pervasive computing (PC), plug-in-based architectures in 239
pervasive computing (PC), service provision approaches in 282
pervasive computing (PC), service-oriented architecture 275
pervasive computing (PC), smart spaces 255
pervasive gaming theory 584
pervasive workflow 2428
Petri nets 636, 641
phenomenological approach 2656, 2657
physical context 236, 993
physical form, definition of 226
PKU part-of-speech tagset 1266
place immediacy 1506
plagiarism 1251
plagiarism detection 1251, 1252, 1254
plan-based problem solving 467
platform for privacy preferences (P3P) project 347, 2606
Index

Plaxo 2681
podcast, definition of 2540
podcastia 2530
podcasts 1322
podcatcher 2540
podcatcher, definition of 2540
pod-people 2530
policy-based mechanism 347
polite computing 2677
politeness, definition of 2677
politeness, information definition of 2679
polychronism/monochronism 132
population analysis, definition of 298
porosity 2451
portability 1511
portals (P) 808, 811, 815, 1307, 1673, 1677
portals (P) and product customisation 1673
portals (P), consumer behavior at 1675
portals (P), definition of 815, 1307, 1677
position-based multicast (PBM) algorithm 1174, 1177
positive e-technologies 2335
postcognitive task models 915
postmodern, technology of 2462
post-visit experience 1639
potential attackers 324
power distance (PDI) 132, 137, 139, 1571
pragmatic habitat 556
predefined actions 2394
predictive modeling algorithms 2172
pre-theoretical research 2309
previous knowledge 1633
pre-visit phase 1642
principle of dialogical reasoning 2096
principle of member verification 2096
principle of multiple interpretations 2096
principles of corporate creativity 2112
principles of universal user experience design 2110
privacy 324, 994, 1496
privacy and security 987
privacy issues 1650, 2606
privacy-enhancing technologies 346
proactive computing 21
problem diagrams 715
problem frames 713, 714
problem representation 379
problem-solving 950
problem-solving oriented learning (PSOL) 951, 954
problem-solving oriented learning (PSOL), cognitive model for 953
problem-solving oriented learning (PSOL), definition of 960
problem-solving oriented learning (PSOL), task ontology for 955
problem-solving process plans 954
process analysis 203
product customisation, portals and 1673
product customization, on the Web 1956
product use 436
productivity 1511
professional development 1505
profiling 214, 810, 812, 815, 1105
profiling, definition of 815
profiling, implicit 214
program manager 2565
programming, end-user 25, 32
project manager 2140
project manager role 2140
project planning 199
project scope management 2139
project sponsor role 2140
project team composition 2139
projection keyboard 271
project-portfolio management 273
proof-of-work (PoW) techniques 349
propositional contents 2083
prosumer 2540, 2600
prosumer, definition of 2540
proxemics 1824
proximity 994, 1976, 1986
proximity tags 353
pseudonyms 1549
psychology 221
public key cryptosystems 341
public key infrastructure (PKI) 327, 341
public key infrastructure (PKI), definition of 331
pull-based service provision 283
push to talk, definition of 1014
push-based service provision 282
pyramid pattern 136
Q
qualitative parameter 797
qualitative study 816
quality of service (QoS) 399, 403
quantitative parameter 797
query refinement 1083
query-based 235
query-based decision aids (QDBA) 1675
question-and-answer speech input 271
questionnaire for user interface satisfaction (QUIS) 363
questionnaires 457
Radio frequency identification (RFID) application fields 1237
radio frequency identification (RFID) privacy 336
radio frequency identification (RFID) security 353
radio frequency identification (RFID) subsystem, integration of 1239
radio frequency identification (RFID) technology 1234
radio frequency identification (RFID) technology, application fields 1237
radio frequency identification (RFID) technology, EuroFlora guide 1241
radio frequency identification (RFID) technology, EuroFlora guide, structure of the interface 1244
radio frequency identification (RFID) technology, integration of RFID subsystem 1239
radio frequency identification (RFID) technology, MADE support of 1237
radio frequency identification (RFID) technology, MADE support, location-aware computing 1237
radio frequency identification (RFID) technology, mobile applications development environment (MADE) architecture 1235, 1238
radio frequency identification (RFID), definition of 1250
radio frequency identification (RFID)-based warehouse management 336
rapid prototyping 642
Rasmussen’s cognitive model 953
Rasmussen’s ladder model, definition of 960
RAW format 2601
RDF(S) 2151
Reactive approach 2167
readiness assessment tests (RAT) 1338
real simple syndication (RSS) 2531, 2532, 2534, 2535, 2536, 2541
real simple syndication (RSS), definition of 2541
real-time enterprises 336, 2144, 2146
real-time process algebra (RTPA) 44, 685, 686, 690, 696
received signal strength indication (RSSI) 1175, 1178, 1179, 1180, 1181, 1182, 1183, 1184, 1185, 1186, 1188, 1189, 1190, 1191, 1192, 1194, 1195
receivers 274
recommender systems 170, 176
reconfigurable context-sensitive middleware (RCSM) 244
reduced function device (RFD) 1039
reference monitor 327
reflexivity 2657
relational closeness 1974, 1987
relational closeness scale (RC) 1994
relational communication 1806, 1816
relationship status 2025
relationship, long-term 1826
relativism 1066
reliability 861, 1962
reliability, tests for 1962
remote access 273
remote control object 286
remote parenting 2662, 2664, 2668, 2669, 2674
remote procedure call (RPC) 242
reporting module 1723
representation of physical action 1820
representational approach 21
requirements analysis 713
requirements definition of 193
requirements engineering 1945
research and development (R&D) 2038
research e-collaboration, definition of 1297
research process 1877
research-based CAL development 1663, 1666
resource centre 1592, 1593
resource constrained speech recognition 1002
resource description framework (RDF) 1097
resource description framework (RDF) 2145, 2150, 2151, 2152, 2160, 2162
resources, allocation of 1050
respect 2675, 2682
responses to attitude 2288
routing 1173, 1174, 1176, 1177, 1178, 1183, 1194, 1196, 1197, 1198
rule-based error 486
rule-based filtering 1630
runtime 636
RUP activities 2094
RUP, integration support in 2093
RUP, structure of 2093
Safety-critical information system 2435
SAGRES system 1634
sales personnel, images of 2235
sales-force automation 273
Index

salutation 1089
SAP implementation, successes and failures of 2214
SAP_HR implementation 2228
SAP_HR implementation, context of 2221
SAP_HR implementation, learning in 2224
SAP_HR project, background of 2221
SAP_HR users, tasks and responsibilities of 2222
satisfaction 1652
satisfaction, definition of 362
scalability 2144, 2145, 2146, 2147
scalability and authenticity 987
scalability, challenges 2145, 2146, 2147
scalable services-based ubiquitous computing, ontologies for 2144
scanners 132, 1503
scatternet 1043
scenario-based design (SBD) 914
scene graph-based frameworks 631
scholarly communication 299
scholarly skywriting, definition of 1297
science museum Web sites, personalization issues 1629
science museums, Web personalization for 1634
screen readers 110
scripting languages 631
search behavior 822
search engines 1677
search engines, definition of 1677
Second Life 1778
Section 508 of the Rehabilitation Act 2543, 2544, 2545, 2551, 2552, 2554, 2555
secure transient association 350
security 323, 1091
security, human-computer interaction and 323
SEGA 1224
self-adaption, definition of 815
self-awareness 1745
self-determination 2326, 2334
self-efficacy 27, 32, 1819
self-enhancement/self-transcendence 139
self-initiated activity 2112
selfish software 2675, 2676
self-regulation 1745
semantic fusion with uncertain inputs 1008
semantic mapping 1103
semantic representation and unification 1007
semantic similarities measurement 746
Semantic Web 1097, 1521, 1522, 1523, 1539, 1541, 1543, 1544, 1545, 1546, 2194, 2201, 2203, 2207, 2208, 2209, 2210, 2212
Semantic Web mining 735
Semantic Web specifications, ontologies and 2148
semantic-level architectures 97
semiotic engineering 449, 450, 451
semiotics 221
seniors 1687
sensor networks 257, 268, 1043
sensor networks, definition of 261
sensors 20
sensory channel permutations 1817
sequential pattern discovery 2172
Serenity 2424
Serenity approach 2430
Serenity framework 2431
Serenity modeling artifacts 2432
Serenity patterns 2432
server 2607
server log 1086
servers, lookup 286
service advertisement 289
service advertisement, definition of 289
service client 281, 289
service client, definition of 289
service discovery database (SDDB) 287
service discovery protocol (SDP) 1038
service federation 2639
service identification, ontologies for 2157
service provider (SP) 281, 289
service provider (SP), definition of 289
service registry 281, 289
service registry, definition of 289
service, composition model 2145
service, definition of 289
service-oriented architecture (SOA) 262, 275, 276, 281, 1098, 2145
service-oriented architecture (SOA), definition of 261
service-oriented architecture (SOA), general view of 281
service-oriented computing (SOC) 280, 280–281, 289
service-oriented computing (SOC), definition of 289
service-oriented context-aware middleware (SOCAM) 244
service-oriented paradigm 288
service-oriented pervasive computing 287
service-oriented smart items middleware, challenges in 2145
service-oriented technologies 286
services, telephony 168
services-based ubiquitous computing applications 2144, 2145, 2160
service-to-device mapping, ontologies for 2159
session object 810, 815
session object, definition of 815
session-persistence engine 276
<table>
<thead>
<tr>
<th>Term</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>sexual harassment</td>
<td>2277</td>
</tr>
<tr>
<td>shared hosting</td>
<td>2610</td>
</tr>
<tr>
<td>shared vision</td>
<td>1505</td>
</tr>
<tr>
<td>short message service (SMS)</td>
<td>269, 1649, 2594</td>
</tr>
<tr>
<td>short message service (SMS) multimodality</td>
<td>274</td>
</tr>
<tr>
<td>short-term orientation</td>
<td>141</td>
</tr>
<tr>
<td>shoulder surfing, definition of</td>
<td>331</td>
</tr>
<tr>
<td>shrinking pipeline</td>
<td>1736</td>
</tr>
<tr>
<td>Silicon Valley</td>
<td>2045</td>
</tr>
<tr>
<td>simple structural criteria</td>
<td>381</td>
</tr>
<tr>
<td>situated action</td>
<td>911</td>
</tr>
<tr>
<td>situated action models</td>
<td>911, 932, 934, 935</td>
</tr>
<tr>
<td>situated interaction</td>
<td>936</td>
</tr>
<tr>
<td>situated interaction paradigm</td>
<td>936</td>
</tr>
<tr>
<td>situated task analysis</td>
<td>915</td>
</tr>
<tr>
<td>situation-adaptable work and information systems modeling method (SIMM)</td>
<td>2399</td>
</tr>
<tr>
<td>situation-adaptable work and information systems modeling method (SIMM) concepts</td>
<td>2400, 2401, 2402</td>
</tr>
<tr>
<td>situational interest</td>
<td>1633</td>
</tr>
<tr>
<td>situational involvement (SI)</td>
<td>2001</td>
</tr>
<tr>
<td>skilled educators</td>
<td>1505</td>
</tr>
<tr>
<td>skills-rules-knowledge model</td>
<td>488</td>
</tr>
<tr>
<td>skimming</td>
<td>353</td>
</tr>
<tr>
<td>skimming, countermeasures against</td>
<td>353</td>
</tr>
<tr>
<td>Skype</td>
<td>2612</td>
</tr>
<tr>
<td>sleep deprivation torture</td>
<td>338</td>
</tr>
<tr>
<td>small group learning</td>
<td>1336</td>
</tr>
<tr>
<td>small wins</td>
<td>1945</td>
</tr>
<tr>
<td>smart environments</td>
<td>2191, 2192, 2193, 2194, 2195, 2196, 2197, 2201, 2203, 2204, 2207</td>
</tr>
<tr>
<td>smart environments, building up</td>
<td>2201</td>
</tr>
<tr>
<td>smart home context</td>
<td>2376</td>
</tr>
<tr>
<td>smart items middleware</td>
<td>2146</td>
</tr>
<tr>
<td>smart items middleware, ontologies for</td>
<td>2157</td>
</tr>
<tr>
<td>smart phones</td>
<td>270</td>
</tr>
<tr>
<td>smart services</td>
<td>2192, 2193, 2195, 2196, 2199, 2200, 2203, 2204, 2207</td>
</tr>
<tr>
<td>smart services, building up</td>
<td>2204</td>
</tr>
<tr>
<td>smart spaces</td>
<td>335</td>
</tr>
<tr>
<td>SNIF-ACT</td>
<td>1160</td>
</tr>
<tr>
<td>sociability</td>
<td>981, 987</td>
</tr>
<tr>
<td>sociability principles</td>
<td>989</td>
</tr>
<tr>
<td>social acceptability</td>
<td>1772</td>
</tr>
<tr>
<td>social action</td>
<td>2077</td>
</tr>
<tr>
<td>social agent</td>
<td>2675, 2676</td>
</tr>
<tr>
<td>social computing</td>
<td>2675</td>
</tr>
<tr>
<td>social computing requirement</td>
<td>2675</td>
</tr>
<tr>
<td>social construction</td>
<td>1760</td>
</tr>
<tr>
<td>social construction</td>
<td>2315</td>
</tr>
<tr>
<td>social construction theory</td>
<td>1754</td>
</tr>
<tr>
<td>social constructivism</td>
<td>1057</td>
</tr>
<tr>
<td>social content</td>
<td>1820</td>
</tr>
<tr>
<td>social context</td>
<td>993</td>
</tr>
<tr>
<td>social cues</td>
<td>1853</td>
</tr>
<tr>
<td>social engineering</td>
<td>325</td>
</tr>
<tr>
<td>social engineering, definition of</td>
<td>331</td>
</tr>
<tr>
<td>social entity</td>
<td>1557</td>
</tr>
<tr>
<td>social factors</td>
<td>1586</td>
</tr>
<tr>
<td>social grounds</td>
<td>2077, 2082</td>
</tr>
<tr>
<td>social group mechanism</td>
<td>1819</td>
</tr>
<tr>
<td>social inclusion</td>
<td>1754</td>
</tr>
<tr>
<td>social influence</td>
<td>148</td>
</tr>
<tr>
<td>social information processing theory (SIP)</td>
<td>1838, 1842, 1845</td>
</tr>
<tr>
<td>social interaction</td>
<td>2078</td>
</tr>
<tr>
<td>social learning model</td>
<td>2066</td>
</tr>
<tr>
<td>social learning, architecture for</td>
<td>2057</td>
</tr>
<tr>
<td>social learning, designing structures for</td>
<td>2061</td>
</tr>
<tr>
<td>social network characteristics</td>
<td>964</td>
</tr>
<tr>
<td>social networking</td>
<td>2037, 2039, 2043, 2046, 2046, 2593</td>
</tr>
<tr>
<td>social networking analysis</td>
<td>1905</td>
</tr>
<tr>
<td>social networking analysis, theory behind</td>
<td>1907</td>
</tr>
<tr>
<td>social networking exercise</td>
<td>1908</td>
</tr>
<tr>
<td>social networking, theories and tools</td>
<td>961</td>
</tr>
<tr>
<td>social networks</td>
<td>646, 649</td>
</tr>
<tr>
<td>social networks of software companies</td>
<td>2043</td>
</tr>
<tr>
<td>social personalization</td>
<td>854</td>
</tr>
<tr>
<td>social politeness</td>
<td>2675</td>
</tr>
<tr>
<td>social presence</td>
<td>2286</td>
</tr>
<tr>
<td>social presence cues</td>
<td>2241</td>
</tr>
<tr>
<td>social presence, definition of</td>
<td>2189</td>
</tr>
<tr>
<td>social purposes</td>
<td>2077</td>
</tr>
<tr>
<td>social relation</td>
<td>2079</td>
</tr>
<tr>
<td>social responsibility</td>
<td>2324</td>
</tr>
<tr>
<td>social roles</td>
<td>2398</td>
</tr>
<tr>
<td>social science</td>
<td>22</td>
</tr>
<tr>
<td>social software</td>
<td>1404</td>
</tr>
<tr>
<td>social software tools, description and classification of</td>
<td>1408</td>
</tr>
<tr>
<td>social software, facilitating e-learning with</td>
<td>1402</td>
</tr>
<tr>
<td>social structure</td>
<td>645, 646, 647, 648, 656, 656, 657</td>
</tr>
<tr>
<td>social world</td>
<td>2077</td>
</tr>
<tr>
<td>social-role theory</td>
<td>1752</td>
</tr>
<tr>
<td>socio-cognitive engineering</td>
<td>677</td>
</tr>
<tr>
<td>socio-cognitive perspective of implementation</td>
<td>1611</td>
</tr>
<tr>
<td>socio-cultural knowledge</td>
<td>2265</td>
</tr>
<tr>
<td>socio-emotional tone</td>
<td>1854</td>
</tr>
<tr>
<td>socio-instrumental action</td>
<td>2077</td>
</tr>
<tr>
<td>socio-instrumental pragmatism (SIP)</td>
<td>2071, 2073, 2077</td>
</tr>
</tbody>
</table>
socio-instrumental pragmatism (SIP)  2082
socio-instrumental pragmatism (SIP)  2390
socio-instrumental pragmatism (SIP) ontology  2391
socio-instrumental pragmatism (SIP), meta-model for  2390
socio-instrumental pragmatism (SIP), meta-model of  2397
sociological determinism  1819
sociological factors  1939
socio-technical approach  980, 983, 986, 988
socio-technical integration  2090
socio-technical perspective  981
socio-technical systems  677, 699, 713
soft security, definition of  1290
soft systems  677
software  25
software agent politeness  2681
software development lifecycles  193, 900
software development process (SDP)  448
software engineering (SE)  198, 433, 448, 449, 450, 631
software engineering lifecycles  199
software implementation  450
software product line (SPL)  760, 761, 763, 764, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 782, 783, 784, 785, 786, 787
software quality assurance (SQA)  2441
software usability measurement inventory (SUMI)  364, 374
software use, increase  2679
software verification  2435
software, component-based  238
space domain indicators  1824
spam  2606
span element  380
spatiality  2265
spatio-temporal relations  1027
speckled computing  2332
speech application language tags (SALT)  275
speech recognition  2559
speech recognition, definition of  2571
speech synthesis, definition of  2571
speech-centric multimodal user interface, definition of  1014
speech-centric user interface design, generic MUI architecture  999
speech-centric user interface design, modality fusion  1006
speech-centric user interface design, special considerations for speech modality  1001
speech-centric user interface design, special considerations for speech modality, context-aware language model  1005
speech-centric user interface design, special considerations for speech modality, modality switching  1004
speech-centric user interface design, special considerations for speech modality, resource constrained speech recognition  1002
spoofing, definition of  331
stakeholders  197, 442
stakeholders integration  2092
StarCatcher  566
status-characteristics theory (SCT)  1752
stereo-based piecewise linear compensation for environments (SPLICE)  1004
stereotypes  1838, 1839, 1840, 1843, 1844, 1848
stereotypes  2031
stereotypes, definition of  1848
sticky, definition of  1307
storytelling  1491
strategy pattern interface  2722
strategy patterns  2722, 2724, 2728, 2730
streaming  403
structural model  1894
student activation  1506
student final exam grade index (STUFIN)  1865
student outcome, definition of  1870
student-centered learning  1307
student-centered learning, definition of  1307
student-centered teaching  1505
subjectivism  1818
subsequent experience  1633
substantive theory  1201
support element  380
support policies  1505
supporter/tender  2029
surprise-explain-reward  26
sustainability  979, 980, 982, 984, 986, 987, 988
sustained management support  2137
Sybil attacks  348
symmetric cryptosystems  340
synchronization  2452
synchronous communication  1865
synchronous communication, definition of  1870
synchronous online communication, definition of  298
synergy  1916
system algebra (SA)  45
system architecture  2604
system communicability analysis  450
system ergonomics  632
Index

system quality  1958
system security  327
system-perspective  632
systems engineering  2430
system’s intelligence, techniques for enhancing  2639
system-specific criteria  631

T

tacit discrimination  2280
tacit knowledge  2040
tag  2594
tag cloud  2594
tailoring  168, 176
tangible bits  22
tangible interfaces  99
tangible user interfaces (TUIs)  630, 1017, 1034, 1388, 1390
task analysis  70, 632, 678
task analysis in HCI  900
task analysis in HCI, approaches to  905
task analysis in Web design  899
task analysis, purpose of  71
task analysis, systemic (STA)  74
task analysis, types of  901
task artefact characteristics chart  1724
task data collection  72
task determinism  1819
task identification  71
task knowledge  793
task model  680
task ontologies  950, 951
task ontology, definition of  960
task representation  72
task selection approach  72
task vs. relationship focus  1816
task-as-activity  903
task-perspective  632
tasks, characterization of  942
tasks, classification of  942
tasks, nature of  930
task-specific criteria  631
taste fabrics  1521, 1523, 1524, 1526, 1527, 1528, 1529, 1530, 1531, 1532, 1534, 1535, 1536, 1537, 1538, 1539, 1540, 1541, 1542, 1543
taxonomies  2106, 2109
Taylorism  70
teachers’ expertise  1511
teaching efficiency  1510
team-based learning (TBL)  1334, 1336
team-based learning (TBL) activities  1337
team-based learning (TBL) framework  1341
team-based learning (TBL) instructional strategy  1339
technical affordances  591
technical assistance  1505
technical integration  2092
techno-economic e-policies  1060
technological advancements  2335
technological determinism  1056, 1057, 1819
technologies, advancement of  262
technologies, input  271
technologists  1594
technologists, definition of  1594
technology  899, 2259, 2260
technology acceptance model (TAM)  29, 132, 133, 142, 1650, 1651, 1655, 1996, 1997, 2287, 2294
technology acceptance model (TAM), definition of  1655
technology attitudes  2277
technology communication networks  1853
technology determinism  1755
technology development  2328
technology features/capabilities  1619
technology mediated learning (TML)  1885
technology mediated learning (TML) learning context  1887
technology trust  2121
technology usage  2289
technology, purpose of  1618
technology, role of  2184
technology-enhanced learning initiatives  1661
technology-induced changes  1620
tech-speak  2286
telecommunication networks  1916
telephone hold time  2404
telework  2023, 2029
test bed, definition of  261
testing activities  198
text domain indicator  1826
text messages  1652
text to speech (TTS) technology  272, 1214–1215
theoretical security  324
theory of proportional representation  1752
theory of reasoned action (TRA)  143, 1997
thread  1823
threaded discussion, definition of  1870
threat modeling  337
time context  236
time of arrival (ToA)  1178, 1179, 1180, 1182, 1184, 1185, 1188, 1189, 1192, 1194, 1195
time orientation  141, 1571
Index

time-division duplex (TDD) 1038
TinyDB 268
TinyOS 268
TinyOS 2713, 2724, 2729
tools-services-content triangle 2324, 2325
topic detection and tracking (TDT) 1679
topic detection and tracking (TDT), definition of 1686
Toshiba EMI 1224
totalitarianism 1066
TouchGlove 635, 641
tourism industry provider 2591
tourist digital assistant (TDA) 1234
tourist guides 168
Toyrs.com 826
trackback, definition of 1297
tracking techniques 1678
tracking, countermeasures against 353
traditional teacher-centered model 1503
traditional teaching 1438
traditional telephones 269
transcription analysis tool (TAT) 1558
transculturalism 1066
transculturality 1061, 1062
transformational learning 1435
transient personalization 169
transmission-control protocol/Internet protocol (TCP/IP) 270
trend 1679
trend detection 1679
trend detection, definition of 1686
trend detection, human factors in development 1678
trend, definition of 1686
triangulation 1188
trilateration 274
triple convergence 2328
Triton 2681
trust 28, 185, 987, 2118, 2141, 2185, 2237
trust, varying dimensionality of 2238
trusted path 339
trusting computers 2435
trusting humans 2435
truth bias (TB) and deception detection 1977, 1987, 1994
turning point, definition of 1686
TVGuide.com 826
two ellipse model 714
typed feature structure, definition of 1014
types of authentication, definition of 331

U
ubiquitous augmented reality (UAR) 629, 630, 640
ubiquitous augmented reality (UAR) systems 636
ubiquitous augmented reality (UAR) user interfaces 629
ubiquitous augmented reality (UAR)-based human-computer interaction techniques 629
ubiquitous commerce 2193
ubiquitous computing (UC) 20, 99, 254, 928, 929, 930, 931, 932, 933, 934, 935, 936, 937, 939, 940, 941, 942, 943, 944, 945, 948, 949, 1015, 1016, 1017, 1022, 1031, 1032, 1033, 1034, 1437–1438, 1443, 2144, 2191, 2192, 2193, 2194, 2198, 2200, 2201, 2204, 2206, 2207, 2209, 2210, 2212, 2213, 2636, 2709, 2710, 2711, 2712, 2713, 2714, 2716, 2722, 2730, 2731, 2732, 2733
ubiquitous computing (UC), applications, characteristics of 930
ubiquitous computing (UC), applications, design of 928, 929, 935, 936, 943, 944
ubiquitous computing (UC), associated challenges 339
ubiquitous computing (UC), associated risks 338
ubiquitous computing (UC), challenges of user interfaces 2634
ubiquitous computing (UC), characteristics 338
ubiquitous computing (UC), definition of 2626
ubiquitous computing (UC), designing for tasks 928
ubiquitous computing (UC), enabling open innovation 2191
ubiquitous computing (UC), environments 2708
ubiquitous computing (UC), human factors 2636
ubiquitous computing (UC), intelligent and dumb interfaces 2637
ubiquitous computing (UC), intelligent user interfaces for 2627
ubiquitous computing (UC), limitations 339
ubiquitous computing (UC), security 336
ubiquitous computing (UC), security for 333
ubiquitous computing (UC), security, taxonomy of 336
ubiquitous computing (UC), systems 2424
ubiquitous computing (UC), tasks in 928
ubiquitous computing (UC), technologies in education 1330
ubiquitous environments 2426
ubiquitous e-technologies environment 2333
ubiquitous information appliances 2331
ubiquitous learning 1330
ubiquitous service computing 262
UK Government Policy and Initiatives 2298
ultrawideband (UWB) 267
UML class diagram 2390
uncertainty avoidance 132, 137, 140, 1571
unconventional orthography 1820
under confidence 32
underground life 2658, 2668
underlying speech technology 2559
underrepresentation of women 1753, 2309
unidimensionality analysis 1961
unidimensionality, tests for 1962
unified communication 272
unified eventity representation (UER) 584
unidimensionality analysis 1961
unification communication 272
unified modelling language (UML) 404–428, 584, 918
uniform resource locator (URL) 269
United Nations Development Programme (UNDP) 124
universal design for instruction (UDI) 2286
universal plug and play (UPnP) 286
universal UXD, principles of 2109
universalism 1066
unknown word guessing (UWG) 1595, 1597
unobtrusive interfaces 1018, 1023
UPi 451, 458
UPi= Test 451, 453, 458
UPnP 1089
urban environments 592, 600
usability 62, 69, 192, 194, 632, 635, 636, 642, 987, 996
usability engineer 192
usability engineering 192, 193, 197, 199
usability engineering 449, 451, 453
usability engineering principles 197
usability hedonics 2109
usability inspections methods 449
Usability Professionals’ Association 2107
usability, definition of 227, 996, 2117
usage data 2169
usage-centered design 2107
usage-centered design, definition of 2117
USB flash-memory music players 1226
use case 1939
use-case decomposition 918
use-case diagrams 407
user 177
user 632, 635, 793
user agent (UA) 3, 4, 6, 7
user attitude 324
user behavior 195
user centred design 432
user characteristics 195
user commanded behaviour frame 714
user context 236
user data 2168
user driven 2166
user engineering 682
user evaluation 383
user expectation 196
user experience design (UXD) 2107
user information needs, patterns of 843
user interaction diagrams 407–428
user interaction frame 714
user interfaces (UI) 69, 194, 196, 394, 403, 404, 448, 929, 936, 944, 2542, 2545, 2546, 2548, 2549, 2550, 2552, 2555, 2627,
user interfaces (UI) adaptation 21
user interfaces (UI) design 1213
user interfaces (UI) design alternatives 449
user interfaces (UI) design patterns 460
user interfaces (UI) design, speech-centric multimodal 997
user interfaces (UI) diagrams 409–428
user interfaces (UI) diagrams generalisation relationships 411–428
user interfaces (UI) diagrams include relationships 411
user interfaces (UI) diagrams, user interface relationships 410–411
user interfaces (UI) evaluation 451
user interfaces (UI) formalization 872
user interfaces (UI) graph 664, 671
user interfaces (UI) management systems 631
user interfaces (UI), building of 2628
user involvement 2141
user involvement and participation 2141
user mobility during the usability evaluation, definition of 2589
user model 395
user modeling 1046
user needs 204
user participation 195, 2141
user participation, Web site design 195
user performance measurement, definition of 2589
user profile 1082, 1630, 2167
user requirements 201
user research 817
user satisfaction 194, 361, 362, 363, 364, 365, 367, 368, 370, 372, 373, 374, 375
user satisfaction measurement, definition of 2589
user search behaviors, patterns of 841
user search strategies, patterns of 842
user types 203
user, interface (UI)
user-adaptive systems 395
user-centered approach 324, 791
Index

user-centered design (UCD)  64, 69, 192, 204, 677, 981, 983, 2107, 2630
user-centered design (UCD) principles  979, 983
user-centered design (UCD), definition of  1014
user-centered guidelines  2110
user-centered interface design  2542
user-centered online learning communities  979, 984
user-centered personalization  180
user-centered personalization measures  180
user-centered processes  192
user-centric design, definition of  1250
user-centric system  951
user-experience design (UXD), definition of  2117
user-interaction models  274
user-interface evaluation  449
user-perspective  632
user’s intelligence, techniques for amplifying  2640
user-specific criteria  631
user-technology dichotomy  2655
user-ubiquitous application interaction dimension  930
user-ubiquitous device dimension  930

V

value-added creation  2112
value-added service  184
variant interactions  768, 773
verbal communication  1548
verbal literacy  1490
verification of service composition, ontologies for  2158
vertical gender segregation  2036
veterinary college  1656
veterinary curricular  1661
veterinary educator shortages  1659
veterinary educators  1656
veterinary information  1658
veterinary medical education  1660
veterinary medicine curricula, HCI in  1656
veterinary profession, globalization of  1657
veterinary teaching hospital challenges  1658
VidaMine  879–894
ViewPad  270
village market pattern  136
virtual chats  458
virtual communities  215, 1380–1381, 2014, 2593
virtual computer  2332
virtual learning communities in higher education  319
virtual learning communities, CMC in  316
virtual network  2014
virtual network computing (VNC), definition of  2589
virtual organizations  2029, 2114
virtual private network (VPN)  273
virtual reality (VR)  2, 4, 6, 7, 630, 1380
virtual sensors  2715, 2718, 2719, 2720, 2721, 2722, 2726, 2727, 2729, 2730, 2731
virtual team  2029
virtual work  1837, 1838, 1839, 1843, 1844, 1845, 1848
visible culture  2330
visioning workshop  442
Visite Plus service  1639
visual data mining  879, 879–898
visual data mining, user interface formalization in  872
visual interface  888–898
visual language  405
visual literacy  1490
visual modeling  404
visual speech recognition  96
visualization  385
Viterbi tagging  1599
vocabulary  1100
Vocera  272
voice gateway  269
voice over Internet protocol (VoIP)  2612
voice technologies, social impact of  117
voice-enabled user interfaces  2556
VoiceXML  269

W

W3C, recommendations  2144, 2145, 2148, 2153, 2160
W3C, semantic web activity  2144, 2148
wasta  125
waterfall development lifecycle  433
wearable computers, computer response to physical activity  1140
wearable computers, emotional impact  1145
wearable computers, finding and retrieving information  1143
wearable computers, human factors  1136
wearable computers, navigation and wayfinding  1142
wearable computers, perceptual impacts  1139
wearable computers, physical effects  1137
wearable computers, reducing size and separating components  1138
wearable computers, supporting memory  1141
wearable computing  22
Web 2.0  1262
Web 2.0 technologies  1404
Web 2.0 tools  968
Web accessibility  2542, 2543, 2544, 2546, 2547,
Index

2550, 2551, 2552, 2553, 2555
Web accessibility, definition of 2555
Web accessibility, guidelines for 2543
Web content mining 1079, 1081
Web design 432, 900, 915
Web design, nature of task analysis in 899
Web design, situated task analysis in 915
Web domain structure 810
Web functionality 819
Web host business provider 2609
Web hosting 2607
Web hosting services 2611
Web information 819
Web interface 212
Web mining 733, 1079, 1086
Web ontology language (OWL) 1098
Web personalization 733, 1086, 2164
Web personalization strategy 850
Web resource 1100
Web server log files 1630
Web service, definition of 2572
Web site analysis and measurement inventory (WAMMI) 364
Web site design 192
Web site design decisions 216
Web site design, user participation in 195
Web site development lifecycles 192
Web site development, user-centered approach to 816
Web site interface design decisions 212
Web site personalization 2358
Web sites 432, 807
Web sites, information-seeking behaviors 816
Web structure mining 1079
Web surfer 1677
Web surfers, definition of 1677
Web tools 1334
Web usage mining 734, 1079, 1086, 1631
Web usage, human factors influencing 1995
Web usage, implications for 2005
Web user interface modeling language (WUIML) 760–790
Web, ontology language (OWL) 2145, 2150, 2151, 2152, 2154, 2155, 2160, 2161, 2162, 2163
Web, ontology language (OWL) DL 2151, 2151–2152
Web, ontology language (OWL) full 2152, 2152–2153
Web, ontology language (OWL) lite 2151, 2151–2153
Web, services 2145, 2161
Web-based computer-mediated communication (CMC) 1334
Web-based information system 791
Web-based media 817
Web-based technologies 1996
WebBoard interface 1338
WebMD.com 826
Web-mining 2174
Webmod task artefact interface 1722
WebQuest tool 2580
Web-supported team-based learning experiences 1337
Weiser, Mark 1, 2, 3, 4, 6, 7, 8, 17
well-oiled machine pattern 136
Wheel of Policy Instruments 702–703
Where Are You Now? (WAYN) 2594
Wi-Fi 231
Wiki engine, definition of 1290
Wiki management, issues in 1287
Wiki page, definition of 1290
Wiki site, definition of 1290
Wiki software, definition of 1290
Wiki, definition of 1290
Wikia 1467
Wikibook wannabes 1271
Wikibookians 1267
Wikibooks 1262
Wikibooks, collaboration process 1269
Wikibooks, future of 1272
Wikimedia Foundation 1467
Wikipedia 1262
Wikipedia 1467
Wikipedia, history of 1264
Wikis 1263–1282, 1318, 1408, 1467
Wikis, nature of 1284
Wikis, tools for collaboration 1283
WikiWikiWeb, definition of 1290
WIM 637
WiMAX 265
WIMP 642
Windows XP 264
wire-based telecommunications infrastructures 2603
wireless access 270
wireless access, laptop 270
wireless federations, ad hoc 114
wireless fidelity (WiFi) network 2591
wireless local area networks (WLANs) 264, 266, 275
wireless local area networks (WLANs) technologies, summary of 265
wireless local loop (WLL) 1593
wireless markup language (WML) 270
wireless metropolitan area networks (WMANs) 265
wireless multi-hop communication 338
Index

wireless NANs 266
wireless networking 20
wireless networking technologies 262
wireless PDAs 270
wireless personal area network (WPAN) 263, 1037, 1043
wireless technologies, high-capacity 169
wireless technology 264
wireless technology, first generation of (1G) 266
wireless technology, fourth generation of (4G) 266
wireless technology, second and a half generation of (2.5G) 266
wireless technology, second generation of (2G) 266
wireless technology, third generation of (3G) 266, 1649
wireless WANs 266, 275
wireless wide area networks, data transmission speed of 267
wireless, growing up 2655
witness learning 317
witness learning, definition of 322
Wizard-of-Oz simulation 940
women and distance education technology 2280
women in decision making 2266
women’s lack of participation 2300
word formation patterns 1597
work interfering with family (WIF) 2025
Workdesk task artefact interface 1722
working memory 2108
working virtually, definition of 1848
World Trade Organization (WTO) 1856
World Usability Day 2107
World Wide Web (WWW) 1097, 2591, 2593, 2602, 2606
World Wide Web Consortium (W3C) 195, 435
world-in-miniature 630, 636
X
XING 1224
XML 1103, 1681
XML documents 892
XML-based communication protocol 890
Y
Yahoo! 171, 2612, 2681
Z
ZigBee 263, 1039
ZigBee, radio-frequency-based 263
zone of proximal development (ZPD) 1464, 1467